Microbiota-dependent reprogramming of the *in vivo* transcriptional profile of *Salmonella* Typhimurium

Von der Fakultät für Lebenswissenschaften

der Technischen Universität Carolo-Wilhelmina zu Braunschweig

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

genehmigte

D i s s e r t a t i o n

von Abilash Chakravarthy Durai Raj
aus Ambasamudram, Tamil Nadu, India

1. Referent: Prof. Petra Dersch
2. Referent: Prof. Karsten Hiller

Eingereicht am: 31.01.2018
Mündliche Prüfung (Disputation) am: 15.05.2018
Druckjahr 2019

Acknowledgement

First and Foremost I would like to express my sincere gratitude to my supervisors, Dr. Till Strowig and Prof. Dr. Marc Erhardt for giving me the opportunity to work in their groups and grow as a research scientist. They gave me freedom to explore the field on my own, at the same time giving me timely advises and support in completing my work. I would like to thank them for their encouragements, constructive criticisms and insightful comments; it is their support and patience that guided me in completing this dissertation. I could not have imagined better mentors for my PhD work.

I would like to thank Prof. Dr. Karsten Hiller and Prof. Petra Dersch for agreeing to be on my examination committee and Prof. Dr. Michael Steinert for serving as the head of my Ph.D. defense commission. I am also very thankful to the HZI Graduate School for the financial support and for the several career development opportunities and soft skills courses provided.

My sincere thanks goes to all the collaborators during my PhD work, especially Dr. Carsten Kroger and Dr. Karsten Hokamp for their valuable suggestions. I would like to thank Prof. Frank Klawonn for the numerous discussions we had regarding my work and providing me timely suggestions in enriching my ideas.

My heartfelt gratitude goes to all the members of the Microbial Immune Regulation (MIKI) and Infection Biology of *Salmonella* (IBIS) research groups in HZI. Without their support, all my research work during these years would be impossible. I am extremely lucky to meet many wonderful colleagues-cum-friends here in Braunschweig. You made my stay in Braunschweig an amazing experience. I would like to thank Dr. Sophie Thiemann, Dr. Till Robin Lesker, Aida Iljazovic, Eric Galvez, Dr. Adrian Blazejewski, Dr.Urmi Roy, Neha Vashisht, Udhayakumar and Nathiana Smit for putting up with an introvert like me, making me feel comfortable and becoming wonderful friends.

Last but not least, I would like to thank my beloved parents Durai Ra and Jeya Rani for all their love and prayers for me. I owe everything to them, who have always encouraged and supported me in every step of my life. I would like to thank my baby brother Gomez Chakravarthy for being my best friend and staying my ‘spiritual guru’.

Summary

Salmonella enterica infection is one of the major causes of food-borne diseases in both developing and developed countries. *Salmonella* possess gene repertoire that allow it to outcompete the inherent commensal bacteria and evade the immune system in the host digestive tract to cause infection. Hence it is highly essential to identify the *Salmonella* gene signatures that could be associated with its infection and gut-microbial signatures that could protect against *Salmonella* infection. In the current study, we introduce a novel operon-based expression profile analysis approach that enabled us to identify functionally coordinated gene signatures characterizing *Salmonella* virulence. We also performed an association study between the gene expression profiles of *Salmonella* and gut microbiota in colonized mouse models. The association study was made feasible by the development of a novel, improved host-specific reference catalogue for colonized mouse models. For further improving the ability of this host-specific gene reference catalogue in identifying the potential microbes and their functional potentials that were highly expressed in the host-gut environment, we devised a novel strategy in this study, by linking the MGS Bins and the 16S rRNA genes. This method highlighted that even a sparse amount of 16S rRNA sequences could be efficiently used to infer taxonomic annotations with improved resolution.

By comparing *Salmonella in vivo* gene expression profiles to *in vitro* growth, we identified that colonization of the intestinal environment as well as competition induces specific expression signatures. We observed that *Salmonella* adjusts its transcriptional machinery in order to compete for nutrients with the host's microbiota, thereby adapting to its environment. Specifically, adaptations to variable availability of nutrients such as mono/disaccharides and co-factors were highlighted by differential expression of uptake and utilization systems and correspondingly, of non-coding small RNAs, which regulate sugar-phosphate stress and carbon-storage activities *in vivo*. Our data supports the observation that *Salmonella* is a metabolically highly flexible enteropathogen employing different transcriptional sub-systems to adapt to the host. We performed a *Salmonella*-gut microbiota association study to identify the microbial gene signatures that could be associated in inhibiting *Salmonella* overgrowth (protection). We observed that the abundance of microbiota-derived nutrients like succinate could be associated with *Salmonella* infection in host gut ecosystem. This study highlights the significance of deciphering the gut microbial complexity in providing protection against enteropathogenic infections.

Zusammenfassung

Eine Hauptursache für Lebensmittelinfektionen und Durchfallerkrankungen weltweit stellt das Pathogen *Salmonella eneterica* dar. Salmonellen besitzten eine Reihe von Genen die es Ihnen erlaubt, die umgebende Darmflora und die Immunantwort des Wirtes zu umgehen und sich erfolgreich im Versauungstrakt zu etablieren. Folglich ist es von essentieller Bedeutung mögliche Gensignaturen zu identifizieren, die für die Infektion von *Salmonella* ursächlich oder mit der Resistenz gegenüber Infektionen durch die Mikrobiota assoziiert sind. In dieser Studie stellen wir eine neue Operon-basierte Expressionsprofil Analyse Methode vor, die es erlaubt funktionell koordinierte Gensignaturen zu identifizieren, die mit der Virulenz von *Salmonella* assoziiert sind. Dafür wurde eine Assoziationsstudie zwischen dem Genexpressionsprofil von Salmonella und der Darmmikrobiota in ausgewählten Mausmodellen durchgeführt, für die ein neuartiger, wirts-spezifischer Genreferenzkatalog für die verwendeten Mausmodelle erstellt wurde. Zur Optimierung dieses spezifischen Genkatalogs zur Identifikation von potentiell wichtigen Bakterien und ihren Eigenschaften wurde ein neues Konzept entwickelt, bei der MGS Bins mit den zugehörigen 16S rRNA Genen verknüpft wurden. Diese Methode erlaubt es auch eine geringe Anzahl an 16S rRNA Sequenzen effizient für die taxonomische Annotierung zu nutzen und somit eine verbesserter Auflösung zu erzielen

Durch den Vergleich von Salmonella Genexpressionsprofilen unter *in vivo* und *in vitro* Wachstumsbedingungen konnte gezeigt werden, dass die Kolonisation sowie die Kompetition mit anderen Darmbakterien dass die Kolonisation sowie die Kompetition mit anderen Darmbakterien im Darm spezifische Expressionssignaturen induzieren. im Darm spezifische Expressionssignaturen induzieren. Salmonella ist in der Lage seine Transkriptionsmaschinerie anzupassen, um mit der Darmmikrobiota um vorhandene Nährstoffe konkurrieren und sich der vorhandenen Umgebung anzupassen. Insbesondere konnte die variable Anpassung zur Verwertung von Mono-/ Disacchariden, Co-Faktoren, und die Regulation von Aufnahme- und Verwertungssysteme durch differenzierte Expression spezifischer kleiner, nicht-kodierender RNAs, die für die Närhstoffspeicherung essentiell sind, unter *in vitro* Bedingungen gezeigt werden. Diese Daten stützen die Einschätzung, dass *Salmonella* ein metabolisch hochflexibler Darmerreger ist, der die Expression verschiedener transkriptionaler Subsysteme nutzt, um sich an spezifische Wirtbedingungen anzupassen. Um mikrobielle Gensignaturen zu identifizieren, die mit einer Inhibition von *Salmonella* Wachstum (Kolonisationsresistenz) assoziiert sind, wurde eine weitere Assoziationsstudie zwischen Salmonella und der Darmmikrobiota durchgeführt. Dabei konnte beobachtet werden, dass die Menge an mikrobiell produzierten, verfügbaren Nährstoffen wie Succinaten mit Infektionen von *Salmonella* im Darm korrelliert. Diese Ergebnisse heben hevor, wie bedeutend neue Ansätze zur Entschlüsselung von komplexen mikrobiellen Gemeinschaften im Darm sind, die einen Schutz gegen Infektionen mit Enteropathogenen gewährleisten.

1 List of Figures

2 List of Tables

3 Abbreviations

TUs	Transcriptional Units
GF	Germ-free
SPF-1	Specific-Pathogen Free-1
SPF-2	Specific-Pathogen Free-2
S. Tm	*Salmonella* enterica serovar Typhimurium SL1344
KO	KEGG Orthology
T3SS	Type-III Secretion System
T1SS	Type-I Secretion System
FPR	False Positive Rate
AUC	Area Under Curve
DEGs	Differentially Expressed Genes
CFUs	Colony Forming Units
WT	Wild Type
PCA	Principal Component Analysis
MGS	Meta Genome Species

4 General Introduction

4.1 Gut microbiota:

The alimentary canal or the Gastro-Intestinal tract (GI tract) is a series of body organs, which enables food intake, digestion, absorption, assimilation and excretion of the undigested and unnecessary food components. The GI tract for food metabolism is common across the monogastric mammals. A proper functioning of the GI tract is strongly associated with the proper functioning of the rest of the systems of the individual since the GI tract is key in providing energy required for the individual to perform daily activities. The GI tract has the large surface exposed to the exterior world (Bengmark, 2013), which puts it in risk of pathogens invading them. Also, the food we intake might possess microbes that try to colonize the nutrient-rich gut region. To counter this, the GI tract is laden with the immune cells, to protect against the colonization of pathogens. Apart from the inherent host digestive and immune system, the gut also harbors a complex microbial ecosystem, which plays a key role in resisting these invading microbes to colonize (Buffie and Pamer, 2013). Gut microbiota is an integral component of the mammalian GI tract whose composition greatly influences the physiology and the health status of the host. They were commonly referred to as the commensal microbiota (Kamada et al., 2013). The gut microbiota constitute 70 % of the microbes residing in the human body (Sartor, 2008); they co-evolve with the host, as they grow and develop a robust intricate network to assist the host in nutrient acquisition, providing protection against pathogens (Kamada et al., 2013) thereby maintaining a homeostatic state. The gut microbiota possess genetic repertoire that encode functions not intrinsic to the host genome. These functions range from degrading complex polysaccharides (Xu et al., 2013), (Chassard and Lacroix, 2013, Flint et al., 2008) ,producing essential fat soluble vitamins-B (Martens et al., 2002) and vitamin-K (Conly and Stein, 1992), to assist host innate immune systems for providing protection against pathogens (Belkaid and Hand, 2014, Thaiss et al., 2016 and Wu and Wu, 2012). One interesting thing to note is that we could observe a drastic variation in gut microbial composition between two healthy organisms (Chung et al., 2012). Many studies (Chung et al., 2012, Nguyen et al., 2015) have indicated a less intra-organism variation in gut microbial compositions when compared to inter-organism

variations. The inter-organism variations could be attributed to difference in relative abundance of different genera between the organisms and the intra-organism variations could be attributed to the difference in relative abundance of different species within each genera between the individuals in the organism. In spite of the species-level difference in the gut microbiota composition amongst the individuals, they possess a microbiota composition whose "functional core" is similar to maintain a homeostatic condition (Turnbaugh and Gordon, 2009). Few studies have also indicated the functional redundancy in the microbial community termed "redundancy hypothesis" (Kang et al., 2015) to ensure the preservation of the significant and essential microbial functions, in contrast to the significance of the presence of few specific "keystone species". Hence, it is vital to characterize the gut microbial compositions of healthy individuals to identify the key functional and microbial signatures responsible to maintain homeostasis.

4.2 Complexity of gut microbiota composition:

The gastrointestinal tract is a highly compartmentalized tract with each section encompassing distinct environment performing distinct and inter-linked roles in food digestion and energy metabolism. Hence, the composition of the microbial communities under homeostatic state varies across the GI tract, right from the oral cavity to the rectum. The gut microbial composition is dependent on the ability of the microbial species to sustain itself under the given habitat and establish symbiotic interactions with the host and other members of the microbial community. The common residual members of the microbial communities include bacteria, archaea and eucarya, whose composition is dependent on the combinatorial effect of the host genetics, food intake and the environmental factors the host lives in. The microbe load in the human gut is not homogenous: with stomach and duodenum harboring 10^1 to 10^3 cells per gram, the ileum and jejenum around 10^4 to 10^7 cells per gram and the highest in colon around 10^{11} to 10^{12} cells per gram (Bull and Plummer, 2014).

4.3 Spatial composition:

In human, esophagus (pH<4) accommodates bacteria of phyla *Bacteroidetes* namely *Prevotella*, *Firmicutes* namely *Streptococcus* as their major inhabitant (Baba et al., 2017). The stomach is highly acidic (pH~2). Proteases (like pepsin) and hydrochloric acid were secreted to digest the masticated food from the mouth to produce chymes. Hence, the microbes present in the stomach should be able to withstand the highly acidic nature. *Helicobacter* is the dominant genus while *Streptococcus* and *Prevotella* (Jandhyala et al., 2015) being other dominant members in the stomach. The

absorption of the digested food from the stomach takes place in small intestine. The small intestine is less acidic with a pH ranging from 5-7. This provides opportunity for many commensal bacteria to inhabit. The common inhabitants of small intestine include bacteria of the phyla *Bacteroidetes* namely *Bacteroides* and of phyla *Firmicutes* namely *Lactobacillus*, *Clostridium, Enterococcus and Gammaproteobacteria* (Jandhyala et al., 2015). The large intestine is largely inhabited by two phyla: *Bacteroidetes* and *Firmicutes* (Rodríguez et al., 2015). In the cecum where breakdown of cellulose and absorption of salts and electrolytes take place, the common active members include *Ruminococcus, Lachnospira, Butyrvibrio etc.* In colon (pH~5.7), the commensal bacteria of species *Enterococcus*, *Ruminococcus*, *Petostreptococcus*, *Faecelibacterium* are the major inhabitants.

4.4 Dysbiotic gut microbiota composition:

When the inherent homeostatic gut microbial composition is disturbed reducing the commensal bacteria, thereby not providing sufficient protection and nutrient competition to the pathogens, it might impact the health status of the individual. In case the individual consume food inhabited by opportunistic pathogens, this might create an imbalance in gut microbiota composition (termed dysbiosis) and could even prove to be detrimental resulting in immune dysregulations like Inflammatory Bowel Disease (IBD) (Zhang and Li, 2014). The common causes of dysbiosis include antibiotics, pyschological stress and diet choices (Hawrelak and Myers, 2004). This dysbiotic state would not only affect the GIT (Gastro-Intestinal Tract), but also other organ systems outside of digestive tract. There were also few studies indicating a strong bidirectional association between gut microbiota composition and development of central and peripheral neural processes (Rhee et al., 2009, Collins and Bercik, 2009) and myocardial systems (Baez and Gordon, 1971, Crawford et al., 2009).

Thus, disturbing a homeostatic gut microbial ecosystem by reducing its microbial diversity and/or density could open doors for pathogens to survive and cause infections.

4.5 Enteropathogenic infections (*Salmonella* pathogenesis)

The pathogens that invade and cause infectious diseases of the gut are called enteropathogens. Enteric pathogens are mostly zoonotic i.e. enteropathogenic infections can be transmitted from animals to humans. These pathogens also evolve and develop strategies to overcome the protection provided by the host immune

system and the microbial composition, by mutations of genomic sequences or by obtaining genetic elements like virulence factors by horizontal gene transfer (Schmidt and Hensel, 2004). These pathogens mainly belong to bacteria of the family *Enterobacteriaceae*, which are Gram-negative and facultative anaerobic bacteria. Even though *Enterobacteriacea* include many harmless common inhabitants of the human gut microbiota (symbionts), performing sugar fermentation and nitrate reduction, they include many harmful bacterium like *Shigella*, *Yersinia, Klebsiella, Salmonella* and *E. coli*.

Salmonella: Amongst the enteropathogens that belong to Enterobacteriaecea family, the genus *Salmonella* is of high clinical relevance across the world and they were the common causes of food-borne illness affecting the young, elderly and individuals with weakened immunity severely. Most healthy individuals affected by the intracellular pathogens *Salmonella* tend to recover without treatment, as they do not enter the blood stream from the intestinal tract. However, if it enters the blood stream and thereby other body sites, it could be life threatening. The common symptoms of *Salmonella* infection (Salmonellosis) include intestinal inflammation (gastroenteritis), fever, and diarrhea within 24 hours of infection. According to a recent report (2017) from the WHO (http://www.who.int/mediacentre/factsheets/fs139/en/), 550 million people fell ill every year because of diarrhoeal diseases and *Salmonella* is one of the four main agents causing such diarrhoeal diseases.

There are two *Salmonella* species that could cause gastroenteritis: *S. bongori* that affect predominantly cold blooded animals (Fookes et al., 2011) and *S. enterica* (~2500 serovars) that affect predominantly warm blooded animals. *Salmonella enterica* infection were communicable, not only between humans but also from animals to humans. In spite of their high genomic similarity with *E.* coli, the horizontal gene transfer system allowed *Salmonella* serovars to acquire genes attributed to virulence and *in vivo* fitness (Bäumler, 1997). Few significant examples include the acquisition of the pathogenicity islands SPI-1 (Mirold et al., 2001) and SPI-2.

S. enterica possess multiple specialized organelles in its armory to modulate host cell functions (Galán, 2001) and to cause infection in a host environment. These include flagellar systems for its movement in the lumen, fimbriae and adhesins to attach itself to host epithelium cells (Fàbrega and Vila, 2013) and injectisomes to deliver effector proteins into the host cells. The injectisomes are of high importance because they help in translocation of bacterial proteins into the host cells. Amongst these systems, there are two significant injectisomes present in the chromosome called *Salmonella* Pathogenicity Islands-1 (SPI-1) and *Salmonella* Pathogenicity Islands-2 (SPI-2). SPI-

1 enables *Salmonella* to interact with the host cell during its initial stage of infection and subsequently mediate in bacterial entry into the non-phagocytic cells, whereas SPI-2 enables *Salmonella* survival within macrophages and subsequently enable systemic infection (Hensel, 2000). Also, few studies have indicated that *S*. enterica has the ability to evolve into multidrug resistant strains (Mølbak et al., 1999). Apart from these functional systems, they also possess plasmids encompassing Type-IV Secretion Systems responsible for conjugal transfer of external genetic components for better *in vivo* fitness of *Salmonella*. These indicate that *Salmonella* genome comprise of genetic elements, that allow it to evolve and to capitalize on any opportunity to survive and cause infection with dysbiotic gut microbial ecosystem. Hence, a complete understanding of the functional potentials and *modus operandi* of *Salmonella* under varied host conditions is highly essential in preventing and treating such *Salmonella* infection outbreaks in the future.

4.6 NGS technologies in studying functional genomics

To study the molecular mechanisms of the invading enteropathogens and the gut microbial community in the host gut environment, the transcriptional landscape of *Salmonella* and the host gut microbial ecosystem needs to be studied. In particular, measuring the abundance of the messenger RNAs that carries the genetic information responsible for the metabolic and regulatory activities is essential to get an insight on the phase of the microbes' life. The latest advancements in high-throughput sequencing has enabled us to study the genome-wide expression profile for microbes with both known genome and uncultured/novel microbes. Microarray and RNA-seq were the two contemporary techniques available to study the activity profile of microbes by measuring the abundances of the RNA transcripts. Microarray quantifies the transcript abundance of a target set of tens of thousands of genes by hybridizing them to their complementary probes embedded in the microarray chips (Schena et al., 1995). However, in case of unstudied genes, the probe sequences were not available for them and hence unable to study them. Also, the abundances of the RNA reads obtained and their quality were quite low (Russo et al., 2003). These were some of the major caveats of microarray technology. RNA-seq allows us to study genome-wide transcriptome by sequencing the cDNA transcripts directly rather than hybridizing them with probe sets (as done in microarray), thereby eliminating major sources of error. The Solexa/Illumina technologies has allowed to sequence ~10^9 transcripts (Mortazavi et al., 2008) sufficient for an entire human transcriptome study. Thus, the advent of Next-Generation Sequencing has given us an opportunity to study genome-wide

expression profile of an entire organism and that of an entire microbial community with much better sequencing depth and less error rate.

4.7 Microbiota-modulated *Salmonella* infection *in vivo*

To develop and improve therapeutics against *Salmonella* infection, it is essential to characterize *Salmonella* virulence in different gut microbial environments and their possible interplay with the host and the other members of the gut microbial community. There have been several studies performed to understand the mode of *Salmonella* infection *in vivo*. A previous study was performed by (Kröger et al., 2013), where they observed the gene expression pattern of *Salmonella* grown *in vitro* in gut infection-relevant conditions like bile shock, anaerobic and aerobic condition etc. Though infection-relevant, *Salmonella* encounter an ensemble of such multiple factors *in vivo* and hence a proper *in vivo* study needs to be performed to understand the potential robust transcriptional machinery and to evaluate the association between such *in vitro* and *in vivo* studies. A recently published study on *Salmonella* Typhimurium by Oshota et al. (Oshota et al., 2017) observed how *Salmonella* reprogrammed its transcriptional machinery under the influence of different infection conditions (host immune responses) *in vivo* based on their transcriptome and proteome expression profiles obtained (6,48, 72 hrs p.i. from the *Salmonella* grown *in vitro* for 16hrs). However, this study did not provide information on the microbial composition under the different host immune conditions.

Another study on *Salmonella* by Spiga et al. (Spiga et al., 2017) observed that microbiota-derived succinate is contributing for *Salmonella* virulence *in vivo*. They observed *Salmonella* utilize succinate in monoassociated gnobiotic mouse models with *Bacteroidetes thetaiotaomicron* (major succinate producers), which enabled *Salmonella* infection. This indicates that microbiota-derived succinate promote *Salmonella* infection. They also observed a complete utilization of the TCA cycle by *Salmonella* during colonization, indicating a strong association between nutrient availability and *Salmonella* virulence. Though significant, these studies do not represent the actual microbial complexity of the gut microbial ecosystem.

Thiemann et al. (Thiemann et al., 2017) observed the influence of varied gut microbial compositions over the severity of disease induced by *Salmonella* Typhimurium SL1344 (*S*. Tm) in a mouse model of acute gastrointestinal disease, thereby characterizing the interplay between pathogen, microbiota and mucosal immune system. They characterized the microbial composition of specific pathogen free (SPF) mouse models – (SPF-1 and SPF-2) before and after antibiotic treatment. Before streptomycin treatment, the species richness quantified by the 16S amplicon sequencing based on

alpha and beta diversity indices, of SPF-1 gut microbial community is lower when compared to that of SPF-2 gut microbial community. After streptomycin treatment, the SPF-1 gut microbial community is enriched with microbes of the family *Deferribacteriaceae* and the SPF-2 gut microbial community is enriched with microbes of the family *Coriobacteriaceae* (Thiemann et al., 2017). They observed that *S*. Tm infection resulted in highest weight reduction and lower survival rate in SPF-1 gut microbial community post streptomycin treatment in comparison to SPF-2 gut microbial community. They also observed that *Salmonella* was able to cause infection both in the lumen and in the cecal tissue under SPF-1 condition, whereas SPF-2 condition was susceptible to *Salmonella* infection in cecal content but provided protection in cecal tissue. This indicates that *S*. Tm infection severity is not correlated to lumen colonization, but on tissue invasion and the protection offered by the SPF-2 microbiota is not by offering a direct competition to nutrients for *Salmonella,* but by limiting its ability to induce tissue inflammation.

Study	Condition	Microbiota	*Salmonella* genome wide transcriptom study
Kröger et.al., 2013	*In vitro*	No	Yes
Oshota et.al., 2017	*In vivo*	No	Yes (Microarray)
Spiga et.al., 2017	*In vivo*	Yes (but on *Bacteroidetes thetaiotaomicron*)	Yes
Brugiroux et.al.,2016	*In vivo*	Yes	No
Thiemann et.al., 201	*In vivo*	Yes	No

Table 1 meta-analysis of the studies of interest

There were also few studies indicating the possible microbial signatures providing protection against *Salmonella* infection in different gut microbial ecosystem. Rangan et al. (Rangan et al., 2016) observed the microbes possessing the gene coding for secreted antigen A (*sagA*) in their repertoire provided protection against *Salmonella* infection in *C. elegans*. Brugiroux et al. (Brugiroux et al., 2016) assembled a minimal set of mouse intestinal bacteria that offers protection against *S*. Tm. They performed a

metagenomics study over such mouse models to understand the functional capabilities that were potentially required to provide protection against *Salmonella*.

The above-mentioned studies provide us information about *Salmonella* infection in many host-associated environments. However, none of them have studied the key association between *Salmonella* infection and a complex microbial ecosystem *in vivo.* In order to enhance the understanding of *Salmonella* virulence *in vivo*, we performed an extensive genome-wide comparative transcriptome study to observe the *modus operandi* of *Salmonella* infection *in vivo* at an early stage, in the presence of varied gut microbial ecosystems.

Choice of experimental model: Experimental models provide us an opportunity to perturb, re-structure and study the association between the host microbiota compositions and pathogen infections. Experimental models allow us to identify the potential microbial agents that facilitate or resist the pathogenic infection at a microbial/molecular level. This is not possible in human studies since they were observational or associative studies (both the cohort studies and the clinical studies); Thus, these experimental model ecosystems allow us to provide an infrastructure to carry out studies in designing novel antimicrobial therapeutics and personalized medicines.

Hence, the choice of the experimental models is of paramount importance. Though many studies were performed on many experimental models, mouse models were widely preferred for the following reasons: a) the convenience in handling the mouse models, b) their genetic identity c) they are one of the well-studied organisms d) their good similarity in physiological processes to human. Hence, the mouse models with different gut microbial complexity provide us an ideal experimental model (Kostic et al., 2013) and thus a tractable model ecosystem to observe and understand the *Salmonella* transcriptional machinery in the presence of gut microbiota with varied complexity for our study.

5 Aim of the study

In this study, we aim in a) Observing and understanding the *Salmonella* transcriptional machinery and their potential interactions with other members of the gut microbial ecosystem that could be associated with *Salmonella* adaptation, virulence and survival in the host GI tract, b) Identifying the resultant microenvironments created by the combinatorial effect of the microbiota and host immune system that could be associated with the protection provided against the enteropathogenic infection. c)

Identifying the gene signatures that provide the *in vivo* fitness for the enteropathogenic infections d). Understanding the relevance in performing *Salmonella* infection-relevant *in vitro* studies, by performing an *in vitro-in vivo* association study.

In order to achieve this aim, we addressed two basic questions: 1. what were the key functional modules crucial for *Salmonella* adaptation in the presence of different gut microbiota compositions in *in vivo* conditions. 2. What were the key micro-environments that could be associated with *Salmonella* adaptation? For our first objective, we organized the *Salmonella* genomic features into protein coding and non-coding RNAs. Since it is more meaningful to observe the expression profile of transcriptional units "as a whole" rather than individual genes, the protein-coding genes were further organized into the Transcriptional Unit (TU) they belong. We then identified those transcriptional units and non-coding RNAs that were differentially expressed under different conditions. Those TUs and non-coding RNAs that gave us an insight on coordinated differential expression and/or regulation were considered to be significant gene signatures. Kröger et al. (Kröger et al., 2013) did a *Salmonella* transcriptome study on 22 different infection-relevant *in vitro* conditions. *Salmonella* encounters a combinatorial effect of these *in vitro* conditions, in addition to the presence of other microbes in the gut region. So, an extensive study on the expression pattern of *Salmonella* Typhimurium in different *in vitro* and *in vivo* conditions is essential to understand the *Salmonella* transcriptional regulatory network better. For our second objective, we also performed a meta-transcriptomic study on the *in vivo* microbial models (SPF-1 and SPF-2) and identified the potential functional and taxonomic signatures that were different amongst these two conditions. Then, we associated the identified meta-transcriptomic KEGG Orthology (KO) signatures to the identified potential *Salmonella* gene-signatures for the potential *Salmonella*-gut microbiota interactions.

6 Methods

6.1 Experimental setup

To study the microbiota-modulated *Salmonella* infection in vivo, the streptomycin-induced *Salmonella* mouse-model with different distinct gut microbial ecosystems were used (Barthel et al., 2003). This experimental model enabled us to understand not only the reprogramming of *Salmonella* gene expression patterns, but also the potential gut microbial signatures associated with *Salmonella* infection.

For this study, C57BL/6J mice with distinct gut microbial ecosystems, namely SPF-1 and SPF-2 (Thiemann et al., 2017), and germ-free mice (without any microbiota) were infected with *Salmonella* enterica spp. serovar Typhimurium SL1344 (*S.* Tm) (with an ampicillin resistance gene) 24 hours after streptomycin treatment (to induce dysbiosis in the microbial ecosystem) and RNA was collected from these samples 12 hours post *Salmonella* infection. Correspondingly, RNA was collected from samples where S. Tm was grown in aerobic and anaerobic conditions *in vitro*.

The mouse models Germ-free, SPF-2 and SPF-1 were treated with Streptomycin by oral gavage. After a day, the streptomycin-treated mouse models were infected with *Salmonella* Typhimurium SL1344 by oral gavage. Then, sampling was done 12 hours post infection. B. *Salmonella* Typhimurium SL1344 was cultured overnight in aerobic and anaerobic conditions. Sampling was performed once we observed an OD value of 0.3.

Figure 6-1 Experimental setup

The collection of RNA from the cecal content allow us to identify the *S.* Tm gene repertoire that could be associated with the initiation and maintenance of the infection. The RNA-seq data thus obtained were mapped against the *S*. Tm genome and the gut microbiota catalogue (Lesker et al. (manuscript under preparation)) to identify the potential gene signatures associated with the *S.* Tm pathogenicity and survival and the corresponding gut microbial signatures.

6.2 Bacterial strains and media.

Cells were grown in either lysogeny broth (LB) (Bertani, 2004) or TB broth (1% Tryptone and 0.5% NaCl). The generalized transducing phage of *Salmonella enterica* servovar Typhimurium P22 *HT105/1 int-201* was used in all transductional crosses (Sanderson and Roth, 1983).

6.3 Mouse models and Salmonella infection

S. Tm strains were grown overnight at 37°C in LB broth, with either 50µg/ml of kanamycin, tetracyclin or chloramphenicol or 100µg/ml of ampicillin. Then this culture was diluted 1:100 in fresh medium, and later sub-cultured for 4 hours. For conducting

the infection experiments, *S.* Tm were washed twice in ice-cold phosphate-buffered saline (PBS).

Each mice was treated with 20mg of streptomycin by oral gavage after withdrawing water and food from them four hours prior. Later, the streptomycin-treated mice were supplied with water and food ad libitum. Twenty hours after the streptomycin treatment, water and food were withdrawn again from the mice, for four hours and then the mice were infected with *S.* Tm, also via oral gavage (10^5 CFUs in 200 µl PBS).

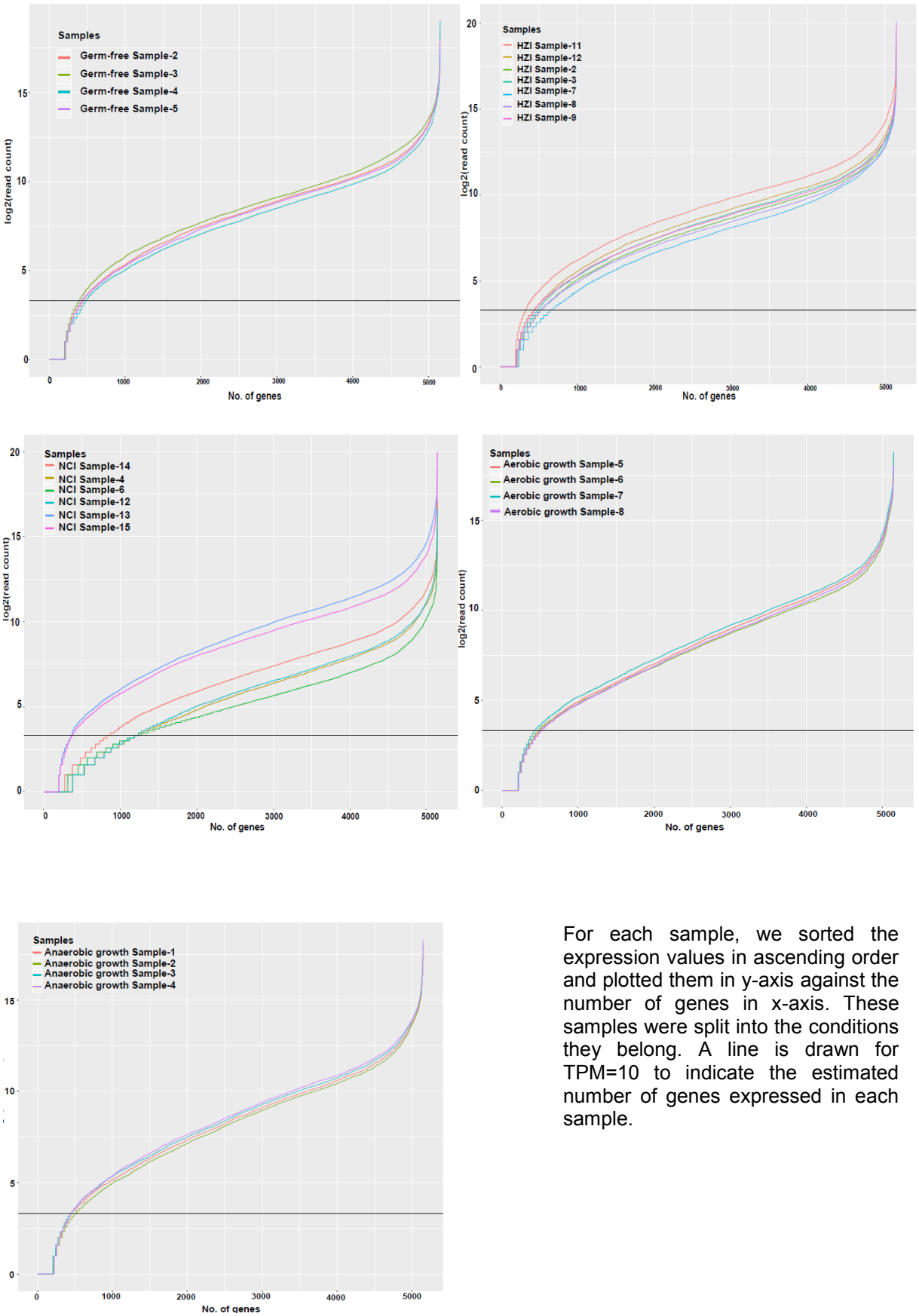
6.4 RNA Isolation and Meta-RNAseq library preparation

High quality RNA was isolated using acid-phenol chloroform based protocol(Faith et al., 2011),(Rey et al., 2010),(Ridaura et al., 2013). Briefly, intestinal content (~100 mg) was immediately preserved using bacterial RNA-protect (AMBION) and cryopreservation. Cell disruption was done using a fastprep (MP) in presence of 200 µL of zyrconia beads 0,1 µm diameter and lysis buffer (200 mM NaCl, 20 mM EDTA), 220µL of 20% SDS, 600 µL of phenol:choloroform:isoamyl alcohol (pH 4.5, 125:24:1, Ambion). After centrifugation at 12,600 x rpm at 4°C, the supernatant was recovered for RNA purification with an equal volume of acid phenol chroform, followed of a centrifugation round. The total RNA was precipitated using 2 volumes of Isopropanol (>3 hours of incubation) and centrifugation for 30 min 12,600 x rpm at 4°C. RNA pellet was washed with 750 uL of cold EtOH 75% and suspended in 90 uL of TE 1X. DNase treatment was done with 2 units of TURBO DNase (Ambion) and the reaction was purified in silica column using RNeasy Kit (Qiagen).

For meta-RNAseq library preparation, RNA quality was evaluated using a bioanalyzer nano-chip (Agilent technologies), samples with a high RNA integrity score (RIN>8.0) were selected for ribosomal RNA depletion (rRNA). Due to the gut microbiota samples contain a high amount of host and bacteria rRNA, we use the Ribo-Zero Gold Epidemiology (Life), this kit allowed us to enrich the total microbiota messenger RNA (mRNA). After rRNA depletion, the mRNA was fragmented to 200 bp using sonication (Covaris) and evaluated again for quality and size. For each sample a total of 100ng of fragmented mRNA were used as an input for cDNA synthesis and Illumina sequencing adaptor ligation. This final part of the protocol was done using a directional RNA Library Prep (NEBNext Ultra), a technology that allowed us to identify the total bacteria gene expression profile (meta-transcriptome) and key RNA-dependent regulatory elements as well.

6.5 Transcriptome-data analysis

Gene count data generation: The forward strand RNA reads were quality controlled using Trimmomatic (Bolger et al., 2014) and the rRNA reads were removed from these reads using sortmeRNA (Kopylova et al., 2012). These reads were then mapped against the mouse genome to remove the mouse-genome related reads. The unmapped reads were mapped against the *Salmonella* Typhimurium SL1344 genome (chromosome NC_016810, Plasmids NC_017718, NC_017719 and NC_017720). The mapped reads were quantified corresponding to the annotation file created by integrating annotation file provided by the NCBI, 280 non-coding RNAs introduced in the Kröger et al. study (Kröger et al., 2013) and the rest from known and predicted non-coding RNAs introduced by Kröger et al. study (Kröger et al., 2012) and generated gene count data. The unmapped reads were then mapped against the gut-microbiota catalogue published by Lesker et al., (manuscript under preparation) and generated the corresponding gene count data. We assessed the quality of the samples by sorting the genes by their expression and observed the expression pattern across all samples.



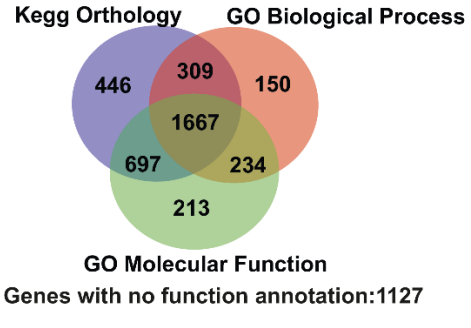
For each sample, we sorted the expression values in ascending order and plotted them in y-axis against the number of genes in x-axis. These samples were split into the conditions they belong. A line is drawn for TPM=10 to indicate the estimated number of genes expressed in each sample.

Figure 6-2 Samples quality assessment

The **Figure 6-2** shows that all the samples in the individual conditions had a similar pattern in expression level for most of the genes and very few genes in the conditions

have small sequencing depth. Hence, all the samples mentioned in the figure were used in the study.

Differential expression analysis: Comparison of the *S*. Tm transcriptome data of two conditions: First, we normalized the data using DESeq2 (Love et al., 2014) and considered genes with fold change ≥2 and p-value≤ 5% to be differentially expressed. The genes were organized into the pathways they belong and identified the significantly enriched pathways for each pair-wise comparison. However, we identified few issues: Not all the genes that were differentially expressed had a KEGG or GO annotation. The venn diagram represent the number of genes with annotations from KEGG and GO databases for *S*. Tm. It revealed that there were at least 1127 out of 4837 protein-coding genes (~24%) do not have any function role assigned by either KEGG or GO databases. In addition, the number of KEGG pathways enriched was too long and most of the enriched pathways did not have a concordant set of differentially expressed genes, i.e. most of the pathways that were enriched did not show a coordinated set of genes that were differentially expressed to indicate an organized activity. Apart from this, with pathway enrichment analysis, identifying the potential gene-signatures is not easy. Hence, we decided to organize genes into the transcriptional units they belong rather than the pathways they were involved. The reason behind this is as follows: a). The genes within a transcriptional units has a common transcription start site and transcription stop site, which indicates they all were expressed at a particular condition in order to carry out a particular function. Hence, identifying transcriptional units that were differentially expressed could give us a clue on the potential sub-functional systems that were relatively active. b). Our results were not limited to genomic regions with KEGG or GO annotations. c). Only 72 genes (~2%) of the overall protein-coding genes did not have any operon annotation for *S*. Tm. This is depicted by the **Figure 6-3**.



A

A. The venn diagram represents the number of genes with KEGG Orthology, Gene Ontology Biological Process, Gene Ontology Molecular Function annotations. B. The bar plot represents the frequency of the number of genes in an operon. C. The bar plot represents the frequency of the number of operons in which every

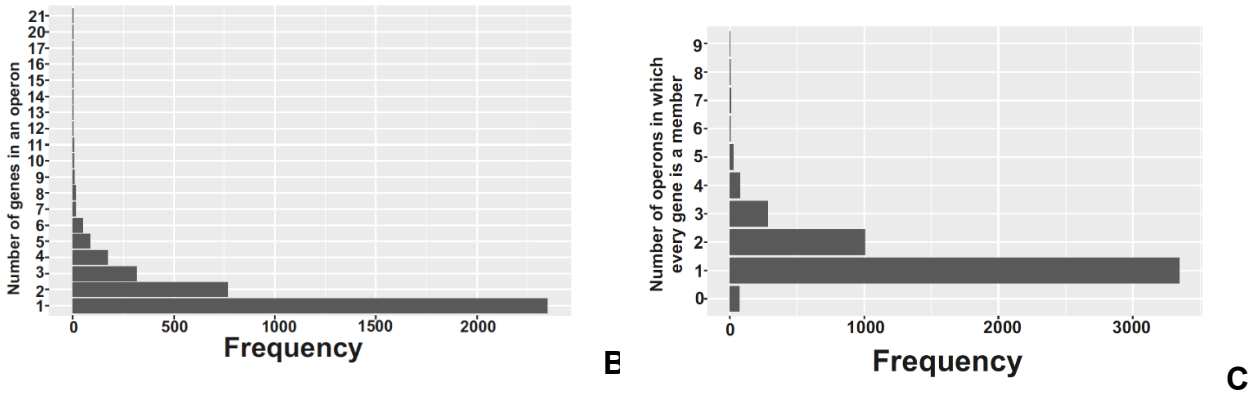


Figure 6-3 *S.* Tm functional annotations and operon structure

Identification of significant Transcriptional Units:

Instead of organizing genes into the known KEGG pathways they belong, we organized the genes into the transcriptional units (TUs) they belong (based on the operons databases DOOR (Mao et al., 2009) and operonDB (Pertea et al., 2009)). We identified the enriched Transcriptional Units based on the assumption that the differential expression or the difference in the transcriptional abundance of the genes in a TU should be consistent across the TU. For this, we obtained the differential expression value of genes for each pair-wise comparison using DESeq2 (Love et al., 2014). Then, we organized the genes into the Transcriptional Units they belong. We obtained the median of the log2Fold Change values (obtained using DESeq2 method) of all the genes in a TU and assigned it as the differential expression value of the TU. Since the p-values of the individual genes in a TU was not independent, we combined the p-values (obtained using DESeq2 method) of the individual genes in a TU using Kost's method (Kost and McDermott, 2002), which was based on the correlation of the TPM normalized genes in TU across the samples in the conditions and considered to provide consistent results for small sample sized RNA-seq studies (Eteleeb, 2015, Eteleeb et al., 2013). Before integrating p-values using Kost's method, we observed the p-value and adjusted p-value distribution obtained using DESeq2 across different comparisons of interest. We expected either a uniformly distributed or anticonservative (with a peak close to zero), but few of the comparisons in adjusted p-value distributions were conservative (with a peak closer to one), especially those for the comparison of *Salmonella* expression profile between Germ-free and SPF-1 conditions and also between Germ-free and SPF-2 conditions. In addition, we observed the combined p-value distribution obtained from both raw p-value and adjusted p-value. Even here, the adjusted p-value distribution showed conservative patterns for Germ-free vs SPF-1

and Germ-free vs SPF-2 comparisons of *Salmonella* expression profile. Based on these data, we preferred to choose the Wald-test based p-value obtained from DESeq2 combined using Kost's method (Kost and McDermott, 2002). The p-value distributions mentioned above were given as **Figure 6-4**:

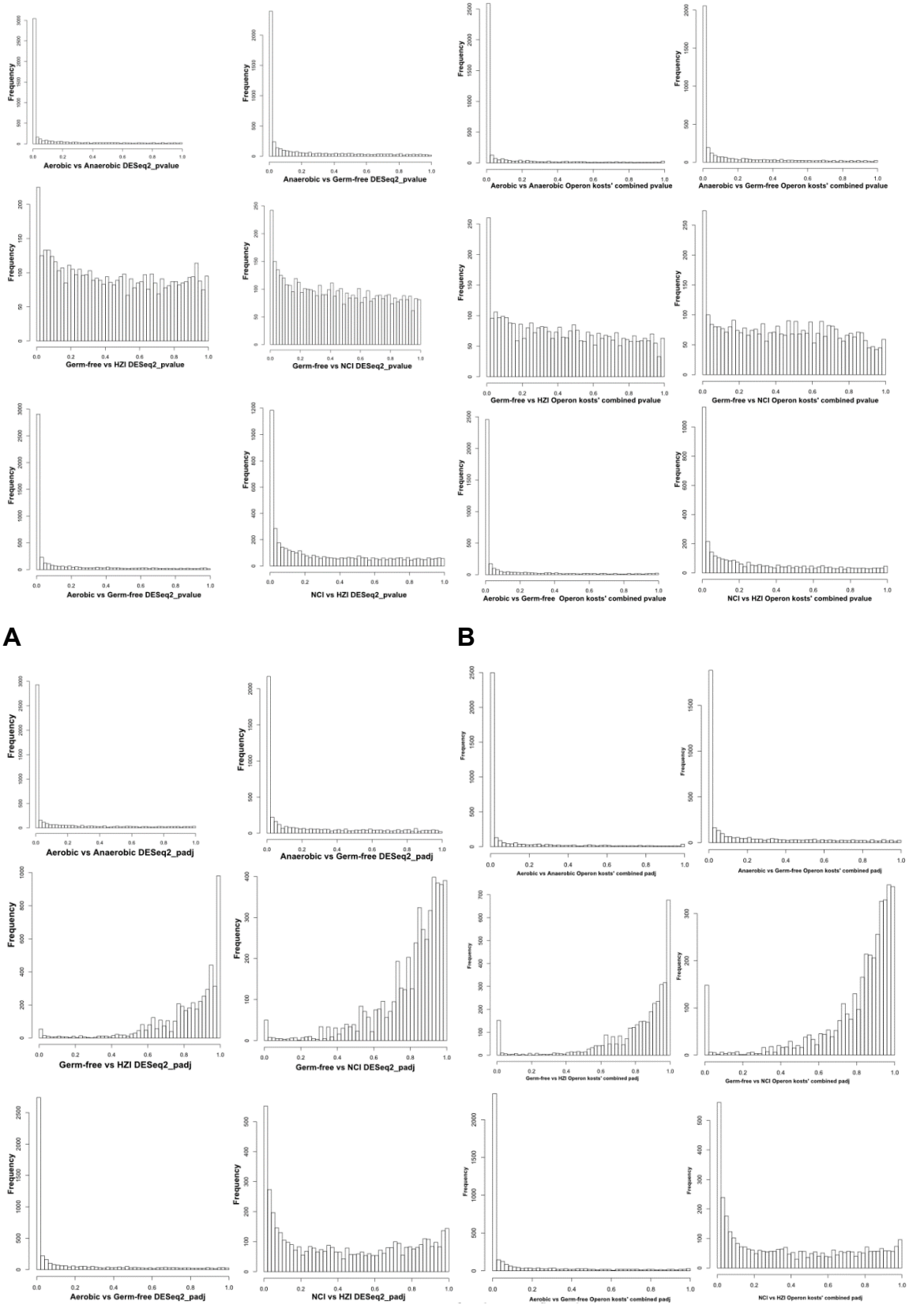


Figure 6-4 p-value distribution of the DESeq2 and operon-model results

These panels represent the histograms for p-value distributions obtained using DESeq2 before and organized genes into operons, before and after multiple hypothesis testing for the p-values obtained DESeq2 for all the pair-wise comparisons in this study. A. The figure represent the p-value distributi genes obtained from DESeq2, the figure-B represent the Kost's p-value distribution of operon figure-C represent the adjusted p-value distribution of DESeq2 genes and the figure-D represen adjusted p-value distribution of Kost's p-value distribution of operon.

Eteleeb et.al identified differentially expressed island regions in each gene (Eteleeb et al., 2013) and then combined them to identify the differentially expressed genes. They used Fisher's method to combine p-values of islands within one gene. We were using the similar concept here, except that we combined the differential expression values of genes within a TU and the corresponding p-values were combined using Kost's method instead of Fisher's method since the genes' expression within a TU is not independent.

The workflow for the operon-based differential expression analysis described is given below.

6.5.1 Workflow

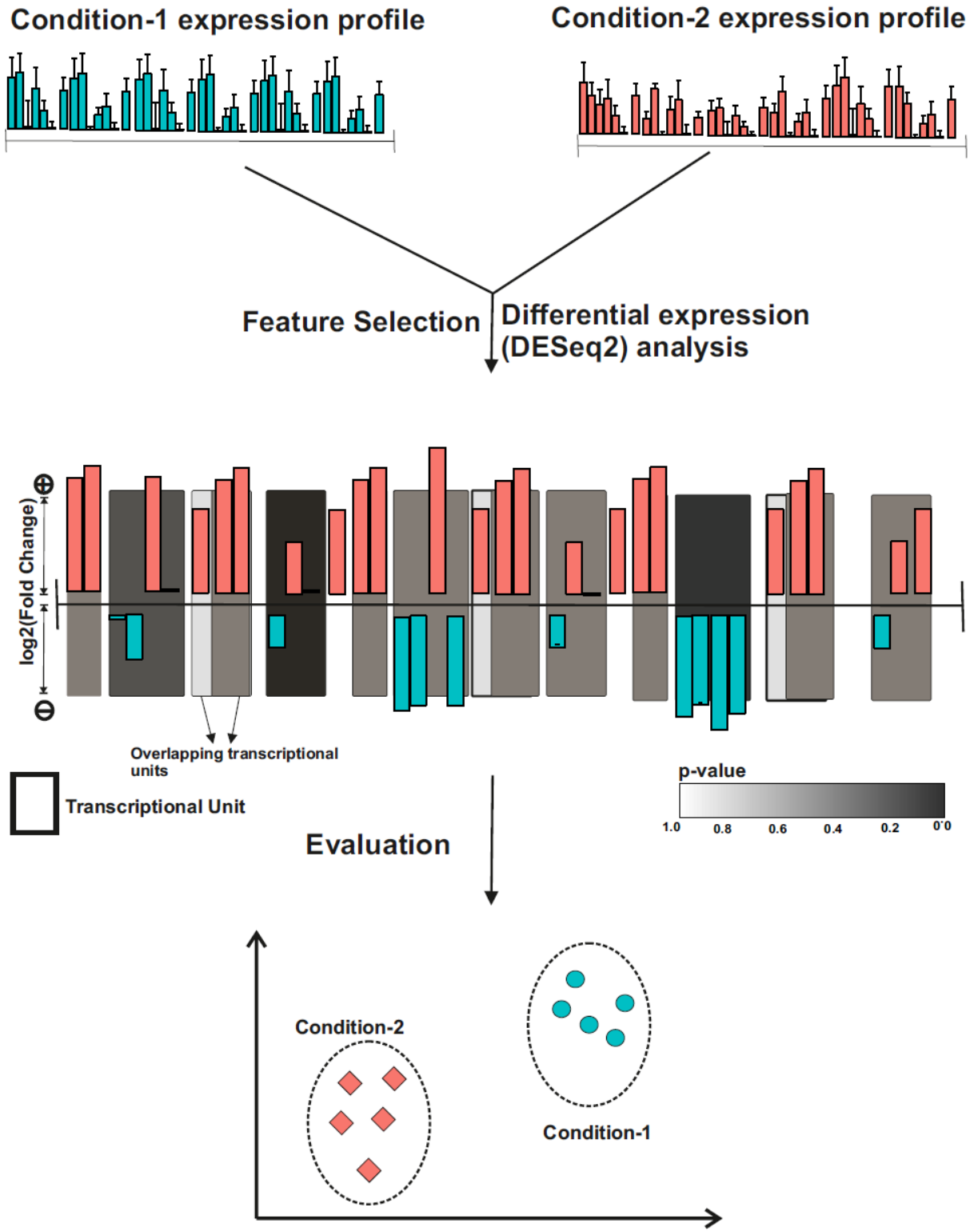


Figure 6-5 The general workflow employed for the protein-coding gene expression analysis

6.5.1.1 Evaluation

The main purposes of the operon-based differential expression analysis proposed in this study include:

1. **Criteria-1:** Identifying those potential gene sets that provide better classification between samples from two conditions (maximize the inter-cluster distances); i.e. to identify the gene sets that represent distinct transcriptional machineries for the two conditions compared.
2. **Criteria-2:** Reduce false positive rates in the differentially expressed genes identified (minimize intra-cluster distances); i.e. to identify the gene sets that represent similar transcriptional machineries for all the conditions in the study.

To assess the performance of the proposed method in comparison to conventional differential gene expression analysis performed, we performed the following evaluation approaches:

Logistic regression of Sammon distance coordinates of the expression data: In this section, we evaluate the efficiency of the proposed method in maximizing inter-cluster distance between the conditions under study (Criteria-1). There were many known metrics available like silhouette index (Rousseeuw, 1987) to evaluate the efficiency of the features selected (operons or genes) to classify the samples from the two conditions effectively. However, the efficiency of these metrics are limited by the fact that most of the prokaryotic transcriptome studies involve small sample size (mostly less than 5). Hence, we devised an evaluation metric using the Sammon distance obtained for the samples under study. Sammon distance allows us to project a high-dimensional data to a low-dimensional data by preserving the inter-samples' distances. It employs a non-linear dimension-reduction approach like principal component analysis (PCA), but unlike PCA method that preserves the overall variance of the data in the lower-dimensions, Sammon distance preserves the inter-samples' distances. We chose using this multi-dimensionality scaling method for cluster evaluation because of the following reasons.

- The multi-dimensionality scaling methods provide us the coordinates that were orthogonal and uncorrelated to each other, hence ideal to be used for classification.
- Since it retains the inter-sample distances, the inter-and intra-cluster distances available on a multi-dimensional scale could be retained in the two-(or three) dimensional scale as well.

Thus, we modelled these orthogonal coordinates of the samples against the conditions they belong via logistic regression, which is similar to PCA regression. We estimated the quality of this model using AIC (Akaike Information Criterion). We compared this AIC score against those models obtained from random but same number of parameters (here operons) of the original data. If the AIC value of the selected features (operons) is smaller than that of the random models, it confirms the ability of the selected set of operons/TUs to better cluster the samples from each condition than any random set of operons.

Scaling the expression data (length-bias in conventional TPM normalization): The dimension reduction methods mentioned above like PCA and Sammon distance could be sensitive to the normalization approaches employed over the original multi-dimensional data (here expression profiles). Hence, the choice of normalization methods of the multi-dimensional transcriptome data is key in obtaining better evaluation of the model. We normalized our transcriptome by organizing the genes into the operons/Transcriptional Units they belong and scaled them using Transcripts Per Million(TPM) method (Wagner et al., 2012) which eliminated the technical biases based on the length of the gene, sequencing depth of the samples at the same time obeying the invariance property more efficiently than RPKM. Though this method was able to remove the bias based on the gene length, in scenarios where all the genes in an operon had a uniform expression pattern with one gene smaller than the other, TPM normalization provided higher expression value to the smaller gene when compared to the others in the same operon. The Operon/TU-based TPM analysis was able to remove this bias since it involve the normalization based on the expression level and length of the operon/TU as a whole rather than the individual genes constituting the operons. The formula for the conventional gene-based TPM normalization is given below:

$$\text{Scaling} = \sum_{i=1}^{N} \text{Gene\_exp}_i \Big/ (\text{Gene\_len}_i - \text{avg. read length} + 1)$$

$$\text{TPM}_{\text{gene}} = \frac{\text{Gene\_exp} \Big/ (\text{OGene}_{\text{len}} - 50 + 1)}{\text{Scaling factor}}$$

Where $Gene_{exp}$ and $Gene_{len}$ represent the gene expression and gene length respectively, avg. read length is subtracted from the gene length to estimate the effective gene length (Pachter, 2011). We replace the gene exp and gene length by the operon expression and length of the operon for our regularized TPM calculation. The TU/operon-based TPM normalization is given below.

Scaling factor= $\sum_{i=1}^{N} {Op\_exp_i}/{(Op\_len_i - avg.read\ length + 1)}$

$$TPM_{op} = \frac{{Op\_exp}/{(Op_{len} - avg.read\ length + 1)}}{Scaling\ factor}$$

Where Op_{exp} and Op_{len} represent the sum of the expression and length of all the genes in an operon respectively and the length of the reads used in the study is subtracted from the length of the operon. To assess whether operon-based TPM normalization works better in removing the length bias within each operon, we observed whether there is any significant difference occur between the expression when normalized across TU/operon and normalized across genes.

To compare the efficiency of organizing the data into the TU/Operons against the conventional gene-based analysis, we normalized each sample by two ways:

1. Method-1: Organize into TU/Operons after TPM normalization, i.e. Conventional TPM normalization on every gene in a sample and then sum the normalized-TPM values of genes in an operon to represent the operon-level TPM value.
2. Method-2: Organize into TU/Operons before TPM normalization, i.e. We sum the raw expression values of the genes in each operon and perform TPM normalization over the operon-expression values (formula given above for TPM_{op}).

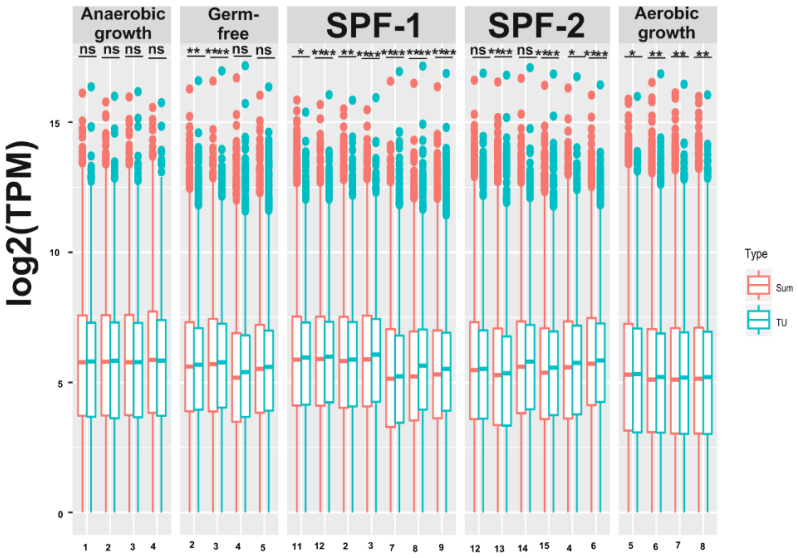


Figure 6-6 Comparative analysis of the operon- and gene-based TPM normalization

The boxplot represents the comparison between Operon/Transcriptional Unit-level TPM normalization (blue boxes) and Sum of the TPM-normalized gene expression belonging to each operon/Transcriptional Unit (red boxes). The asterisks represent the statistical significance for the comparison between the two normalization strategies for each operon in a sample.

Symbol	pvalue
Ns	Not significant
*	pval<5%
**	pval<1%
***	pval<0.001
****	pval<0.0001

Table 2 pvalue symbols

We identified a significant difference in the expression levels of operons when normalized by the two different methods mentioned above (as shown above in the **Figure 6-6**). To observe whether there is any influence of the length of the gene in each operon when normalized by method-1, we modelled the z-normalized length of the genes in each operon against the expression levels in each sample individually. The formula for this process is given below:

$$\mathrm{Z\_length}_{\mathrm{gene-1,operon-1}} = \left(\mathrm{gene}_{\mathrm{len}_1} - \mathrm{mean}(\mathrm{gene}_{\mathrm{len}})\right) \Big/ \mathrm{standard\_deviation}(\mathrm{gene}_{\mathrm{len}})$$

$$\mathrm{Z}_{\mathrm{length}_{\mathrm{genes\ in\ operon-1}}} \sim \mathrm{expression\ in\ sample1, condition1}_{\mathrm{genes\ in\ operon1}}$$

Where $Z_{length_{gnees\ in\ operon-1}}$ is normalized length of all the genes in an operon and $gene_{len}$ is the length of the genes in an operon. Based on this method, we observed in all the samples across the conditions, atleast 50 operons possessing genes of varied length has an influence of the length of the genes in the conventional TPM normalization. In addition, most of the operons showed this effect had 3 or 4 genes with few possessing more than 10 genes. The landscape of the distribution of the number of genes that belong to such operons could be inferred from the **Figure 6-7**.

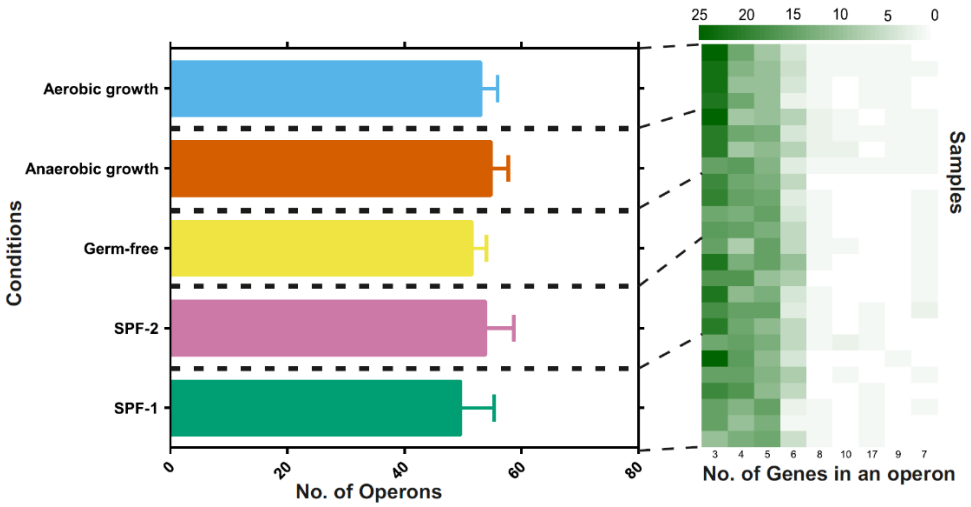


Figure 6-7 Influence of gene length over the gene-based TPM normalization

The barplot represent the number of operons that were identified to have an influence of the length of the genes over their TPM-normalized expression values. The heatmap represents the number of such operons possessing a certain number of genes. The x-axis of the heatmap represent the number of genes in these operons and y-axis represent the samples. The samples belonging to each condition were marked to the bars representing the corresponding conditions.

This shows us there is an influence of the length of the genes greatly in small operons (possessing 3 or 4 genes). The normalization based on method-2 do not have the influence of the length of the genes in an operon since they were normalized based on the length of the operon overall, from the transcription start site to the transcription stop site.

We were interested in observing the efficiency of the methodology mentioned above in classifying samples for multiple conditions efficiently using other ordination strategies like PCA. For this, we picked all the samples for the in house *in vivo* conditions. We obtained all the differentially expressed *Salmonella* operons obtained (Fold Change >2 and Kost's p-value <5%) for Germ-free, SPF-1 and SPF-2 conditions when compared against each other and created an ordination plot (here PCA) for the samples from such differentially expressed operons. The PCA for such comparison is given below. There were 623 operons identified to be differentially expressed for atleast one of the pair-wise comparisons. It possessed around 71% of the overall variance of the data and an AIC score of 22.6. Any other random set of operons of length 623 were neither able to provide a better classification of the data (only 0.33% i.e. 33 out of 1e+04 random selection of 623 operons were able to provide similar classification efficiency of AIC 22.6) not able to recover variance 71% or more.

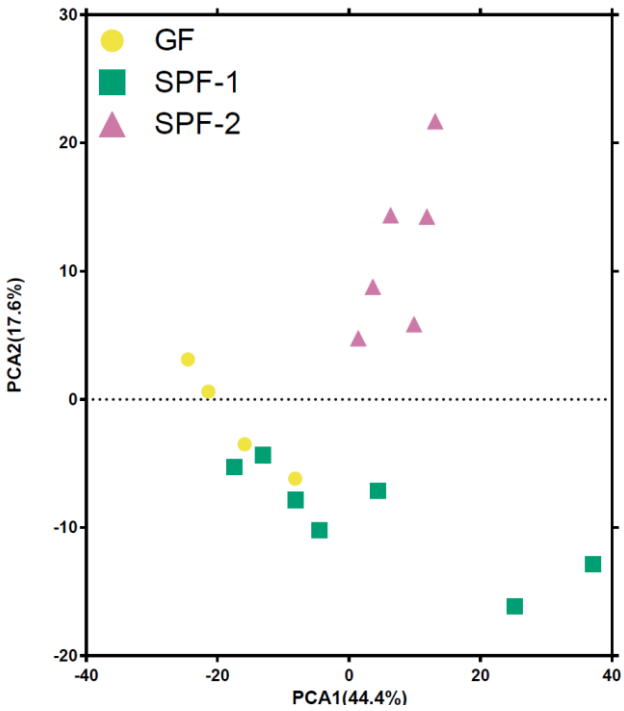


Figure 6-8 PCA plot for the *in vitro* and *in vivo S*. Tm transcriptome samples

We obtained all the differentially expressed *Salmonella* operons obtained (Fold Change >2 and Kost's p-value <5%) for Germ-free, SPF-1 and SPF-2 conditions when compared against each other. We obtained the ordination plot for all the *in vivo* S. Tm gene expression profiles based on these differentially-expressed. The x-axis represents the Principal Component-1 and y-axis represents the Principal Component-2 constituting 44.4% and 17.6% of the overall variance respectively.

These analysis strategies indicate that organizing the gene expression profile into operons/TUs they belong could greatly reduce both the false positive and false negative rates and evaluation methods indicate that this analysis strategy allow us to maximize the distance between samples from two different conditions. Now, in the next section, we evaluate the method's robustness in minimizing false positive rates.

6.5.1.1.1 Reduction in False Positive Rates (Intra-group differential expression):

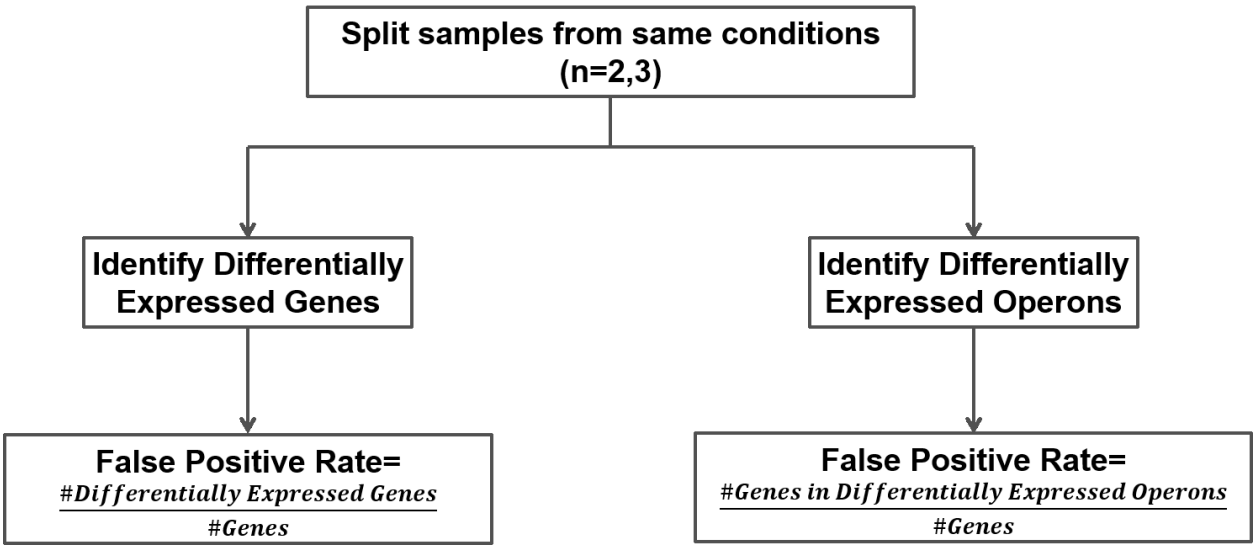
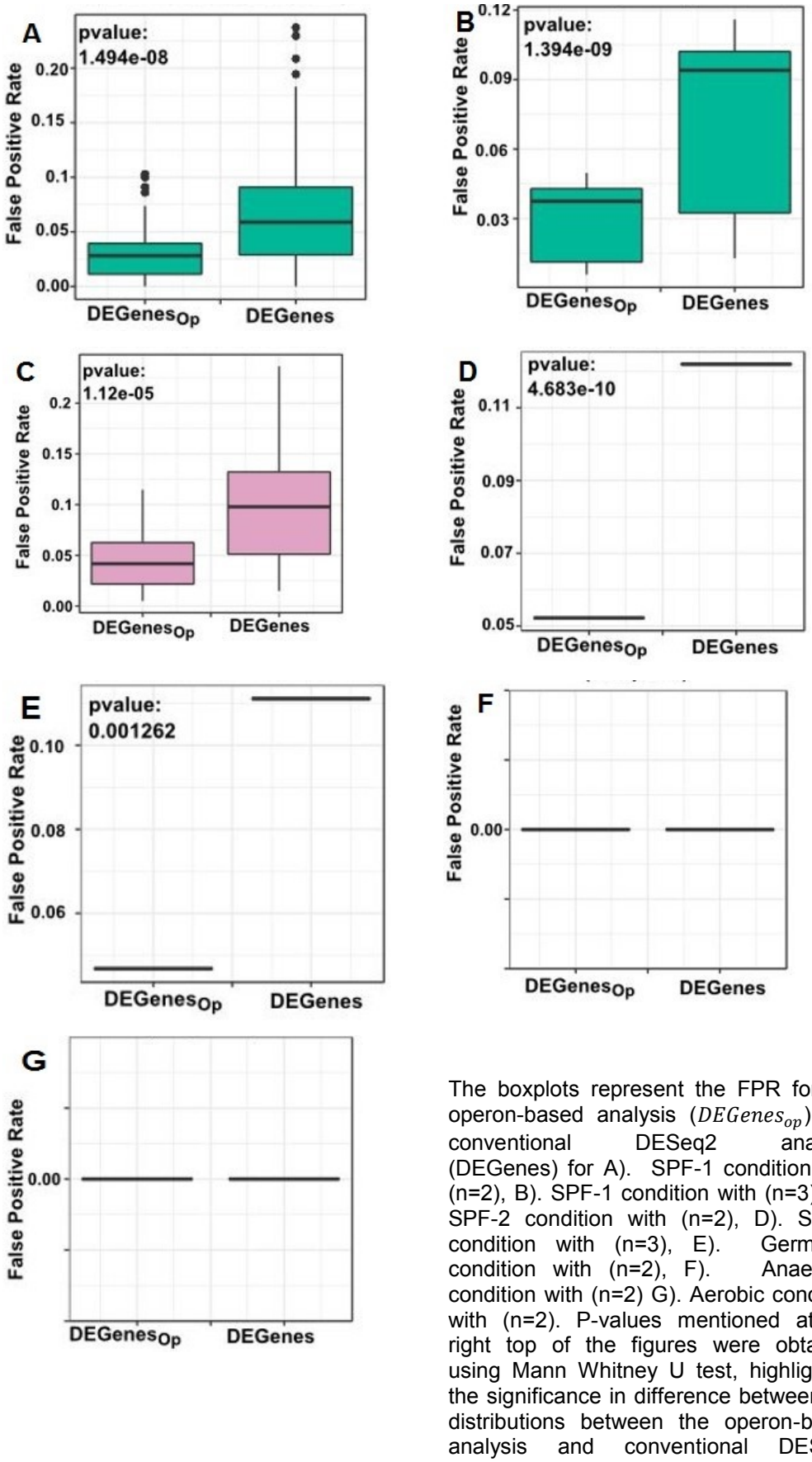


Figure 6-9 Workflow for estimating the false positive rates for the methods involved in obtaining differentially expressed genes and differentially expressed operons.

It is a common assumption that the samples obtained from the same situation should not possess different expression pattern. In other words, the intra-group variation should be small. Hence, any gene that shows a differential expression (fold change >2 and p-value <5%) when two random sample sets drawn from samples of same condition were considered a false positive. We quantified the number of differentially expression genes obtained for all possible random sample sets drawn from samples of same condition. We repeated this for both the operon-based analysis and conventional DESeq2 gene analysis for all the conditions. It is depicted in the **Figure 6-9**. In all the conditions, the operon-based analysis were able to reduce the false positive rates with statistical significance (Mann Whitney U test) in the *in vivo* conditions. The *in vitro* conditions did not have any gene with a fold change greater than 2 and p-value <5%. From this, one could infer that analysis on *in vivo* samples were highly prone to false positive results and identifying differentially expressed operon was able to reduce the false positive rates effectively in comparison to that of differentially expressed gene. This is depicted in the boxplots shown below.



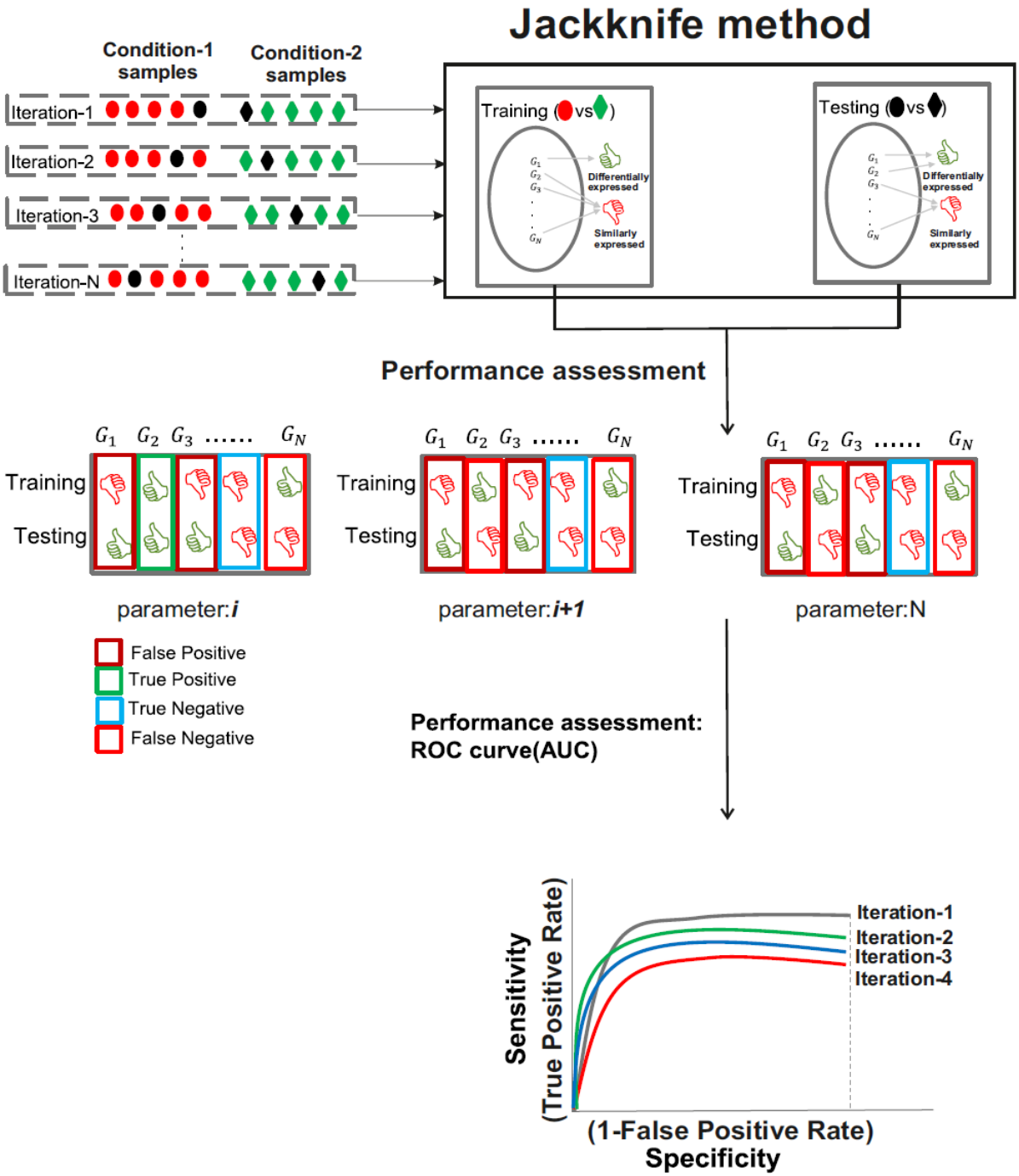
The boxplots represent the FPR for the operon-based analysis ($DEGenes_{op}$) and conventional DESeq2 analysis (DEGenes) for A). SPF-1 condition with (n=2), B). SPF-1 condition with (n=3), C). SPF-2 condition with (n=2), D). SPF-2 condition with (n=3), E). Germ-free condition with (n=2), F). Anaerobic condition with (n=2) G). Aerobic condition with (n=2). P-values mentioned at the right top of the figures were obtained using Mann Whitney U test, highlighting the significance in difference between the distributions between the operon-based analysis and conventional DESeq2 analysis.

Figure 6-10 False Positive Rates (FPR) based gene-based and operon-based analysis

From this, we could infer that intra-cluster distance could be reduced by organizing genes in to the operons/TUs they belong. In the next section, we assess the performance of the method.

6.5.1.1.2 Performance of the method

We were interested to assess whether organizing the expression data into operons/TUs would provide us a reproducible set of genes differentially expressed, even when we possess only one sample, and thereby assessing the performance of the method in providing better classification between samples belonging to two different conditions. For this, we employed jackknife-resampling strategy. To briefly put, we randomly picked all but one sample from two conditions and used these samples to identify differentially expressed operons (Fold-Change >2 and p-value <5%) and segregated *Salmonella* operons into those that were differentially expressed and those not. We performed the same operation using the left-out sample from both the conditions. We compared the results in both the comparisons by calculating the area under curve (AUC) for these comparisons. An AUC close to 1 means, even a comparison between two conditions with one sample each would provide similar result, to scenarios where we possessed more samples on both sides. We repeated this for all possible combinations of samples, to assess the ability of the method to produce similarity in identifying the differentially expressed genes. We compared the AUC obtained from the operon-based method to that of the conventional gene-based differential expression analysis method. A diagrammatic representation of this method is given below in **Figure 6-11**.



For each pair-wise comparisons, we left one sample from both conditions (training data), obtained the differentially expressed genes/operons (Fold Change >2 and p-value<5%). Simultaneously, we obtained the differentially expressed genes/operons by comparing the left-alone samples (testing data). We compared the results obtained from training data and testing data by obtaining their AUC values.

Figure 6-11 Workflow for the jackknife method employed.

We obtained AUC values for these comparisons for both operon-based and gene-based methods. The boxplot below represent the range of these AUC values. The **Figure 6-12** shows that

1. The Operon-based method provide significant (indicated by the asterisks) and better mean AUC values for all the pair-wise comparison of the conditions to that of the AUC values provided by the conventional gene-based methods.
2. The Operon-based method provide AUC values with smaller range in comparison to that of the gene-based method. This means, the method is robust enough to provide reduced false positives if we compared one random sample each from two conditions.

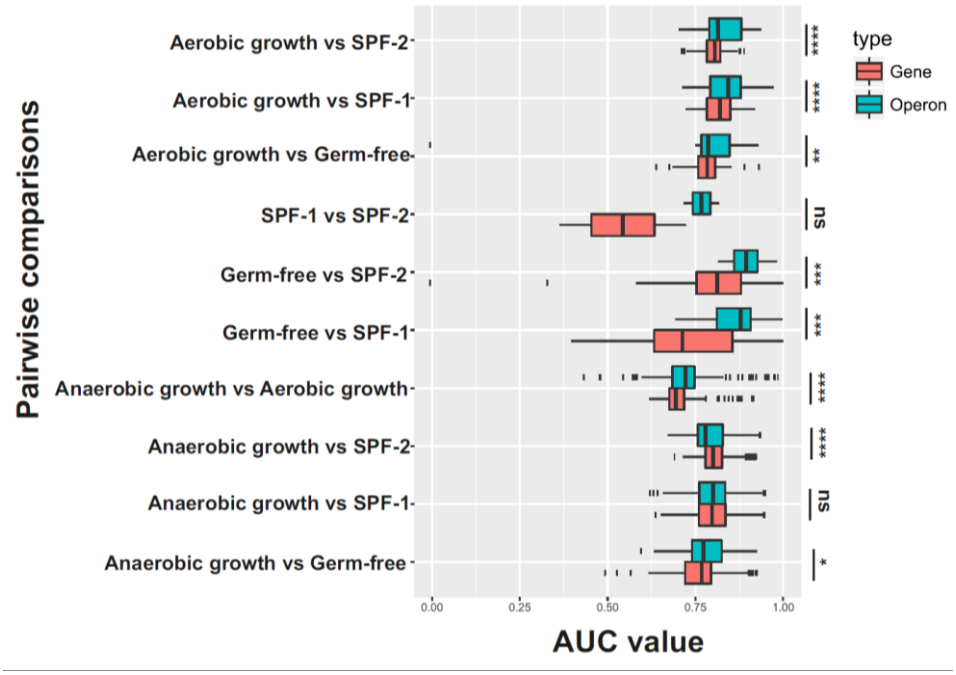


Figure 6-12 Performance assessment of the gene- and operon-based analysis by comparing AUC values

The boxplot represents the area under the curve obtained from both conventional gene-based and operon-based differential expression analysis for all possible combinations of samples for all pair-wise comparisons in this study. The blue boxes represent the AUC values obtained from operon-based differential expression analysis and the red boxes represent the AUC values obtained from conventional gene-based differential expression analysis. The asterisks indicate the statistical significance calculated based on Mann-Whitney U test.

Also, the operon-based method identifies a slightly reduced number of genes to be differentially expressed in comparison to the conventional DESeq2 analysis, which might be a major factor contributing to the reduced false positive rates. This is depicted in **Figure 6-13**.

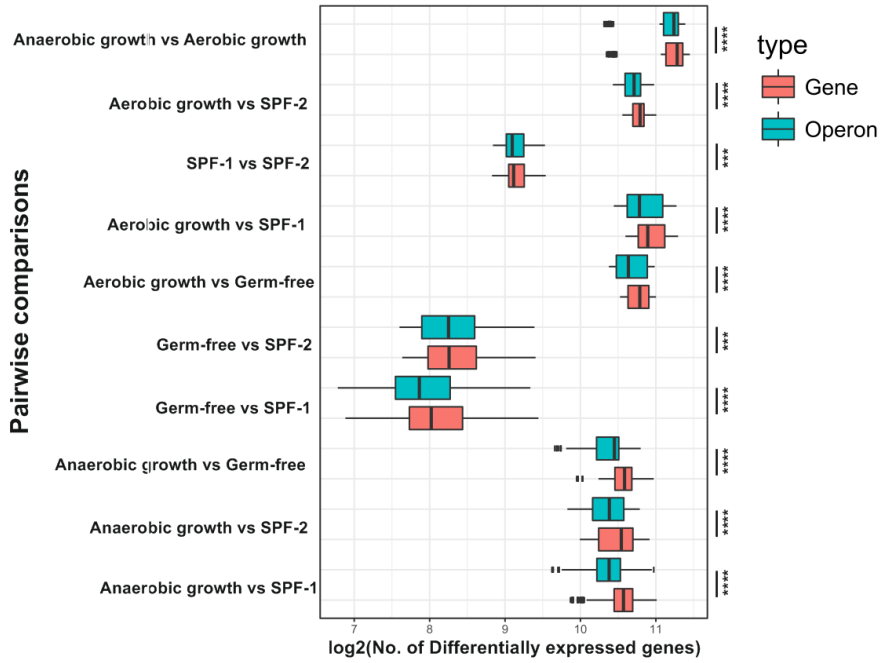


Figure 6-13 Performance assessment of the gene- and operon-based analysis by

The boxplot represents the log-number of the genes differentially expressed obtained from both gene and operon-based methods from all all-possible combinations of samples for all pair-wise comparisons in this study. The blue boxes represent the AUC values obtained from operon-based differential expression analysis and the red boxes represent the AUC values obtained from gene-based differential expression analysis. The asterisks indicate the statistical significance calculated based on Mann-Whitney U test.

comparing number of DEGs for random combination of samples

6.5.1.1.3 External evaluation

For external evaluation, the *E.*coli data used by Rockhopper tool (McClure et al., 2013) to evaluate their method is used here as well. This data possessed the differential expression value of selected set of genes and their corresponding RT-qPCR validation. The compilation of this is presented below in the table.

GeneID	Gene Names	Rock_hopper RNA-seq	Rock_hopper qt-PCR	operon_log2fc	operon_padj
b0726,b0727,b0728,b0729	sucC	plus	plus	2,25	1,24E-24
b0727,b0728,b0729	sucC	plus	plus	2,28	8,99E-27
b0726,b0727,b0728,b0729 b0721,b0722,b0723,b0724	sucC	plus	plus	1,82	9,26E-16
b1729	ydjN	neg	neg	-2,23	5,23E-19
b2597	raiA	plus	plus	1,10	7,08E-07
b2964	nupG	plus	plus	1,95	6,13E-10
b3748,b3749,b3750,b3751	rbsD	neg	plus	-0,75	0,003389369
b1378	ydbK	plus	plus	1,76	6,05E-10
b2091,b2092,b2093,b2094 b2095,b2096	gatZ	plus	plus	2,11	1,92E-08
b2276,b2277,b2278,b2279	nuoE	plus	plus	1,03	3,00E-05

b2280,b2281,b2282,b2283 b2284,b2285,b2286,b2287 b2288					
b3686,b3687	ibpA	neg	plus	-2,72	1,03E-40
b3962	sthA	plus	neg	1,34	4,56E-10
b3748,b3749,b3750,b3751 b3752,b3753	rbsD	neg	plus	-0,41	0,013355488

Table 3 Operon-based enrichment analysis in *E*.coli data

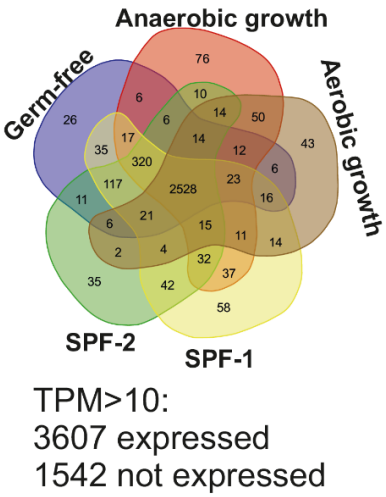
Briefly, plus in Rockhopper RNA-seq and RT-qPCR indicate the corresponding gene is highly expressed when grown in LB medium with alpha-MG condition and neg in Rockhopper RNAseq and RT-qPCR indicate the corresponding gene is highly expressed when in LB medium without alpha-MG. Keeping RT-qPCR as reference, the RNA-seq agreed most of the time except for *rbsD,sthA,ibpA* genes. We employed our method to observe the efficiency of our method in improving the false positive and false negative rate. By our method, we proposed the operon containing the *rbsD* gene did not have fold change >2 with a Kost's p-value <1%. Thus, we reduced the false positive rate by rejecting this gene to be differentially expressed. However, we were not able to identify the other false positives from our method because 1) *sthA* gene is a single-gene operon and hence Kost's p-value integration and fold change mean cannot be performed.

6.6 Analysis of bacterial loads in intestinal content and systemic organs

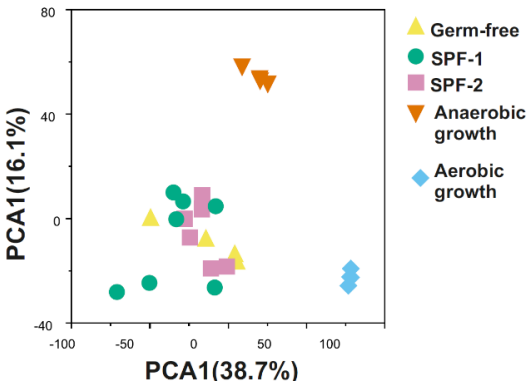
In order to evaluate the importance of the operons/Transcriptional Units (TUs) that could be associated with the *S*. Tm infection, competition experiments were performed. Using the same methods to grow the *S*. Tm and treating the mice as mentioned earlier, the different strains (mutant of the potential gene signatures) were all diluted to 10^5 CFU per 200 µl PBS and the mice were infected with three mutants and the wild type simultaneously. Euthanization of all the mice was done, by asphyxiation with $CO_2$and cervical dislocation at the indicated time points. Intestinal tissues from small intestine, cecum, and colon were removed. The organs were opened longitudinally, cleaned using PBS and weighted. Homogenization of the organs and contents were done in PBS using Polytron homogenizer (Kinemtatica) and these homogenized samples were diluted. The diluted homogenized samples were plated on LB plates containing 50 µg/ml of kanamycin. The *S*. Tm CFUs were determined after incubation for 24 hours at 37 degree Celsius.

7 Results

First, we performed a qualitative comparative analysis by comparing the number of genomic features that were expressed/not expressed across all the conditions (irrespective of their scale of expression). This is an indication on what fraction of genomic features is transcribed in a particular condition irrespective of their scale of expression. We considered those genes with a median TPM value greater than 10 to be active. We observed that there were 2528 out of 5149 genomic features i.e. ~49% of the *S*. Tm genomic features active in all the conditions. 117 out of 5149 genomic features i.e. ~2% to be active in only the *in vivo* conditions and 50 genomic features (~1%) to be active in only the *in vitro* conditions. In addition, 1542 genomic features i.e. ~30% of the *S*. Tm genomic features were not active in any of the conditions under study here. This analysis gave us an indication that not all the genomic features were transcribed under all conditions, and it activate or repress a small set of genomic features under specific condition. Hence, identifying these genomic features associated with the conditions under study is essential to enhance our understanding on the *S*. Tm transcriptional machinery.



A



B

Figure 7-1 *S*. Tm transcriptional machinery *in vivo* and *in vitro*

A. The venn diagram compare the estimated number of genes expressed in individual conditions (both *in vitro* and *in vivo*) to each other. For this, we obtained the average TPM normalized expression of a gene in each condition. We estimated the number of genes expressed by the assumption that a gene with TPM>10 to be expressed as discussed in Kröger et.al 2013, and FANTOM consortium. B. PCA plot of gene expression profile of all the *S*. Tm samples: The within-sample normalization of Salmonella expression profile was performed using Transcripts Per Million (TPM) method and PCA plot was obtained from these normalized data. Here, PCA1 represent 38.69% of the overall variance and PCA2 represent 16.1% of the overall variance.

The PCA plot in **Figure 7-1**-**B** indicated that the *S*. Tm expression profile was distinct for the *in vitro* conditions, whereas it had comparatively similar expression pattern in the *in vivo* conditions. This shows that in the gut region of the *in vivo* models, *S*. Tm has a robust transcriptional machinery that responds differently to only few specific micro-environments that either provide challenge to survive or opportunity to cause virulence; in other words, *S*. Tm make few specific modifications in the transcriptional machinery, but not having a drastic change in their transcriptional machinery in different *in vivo* gut microbiota environment.

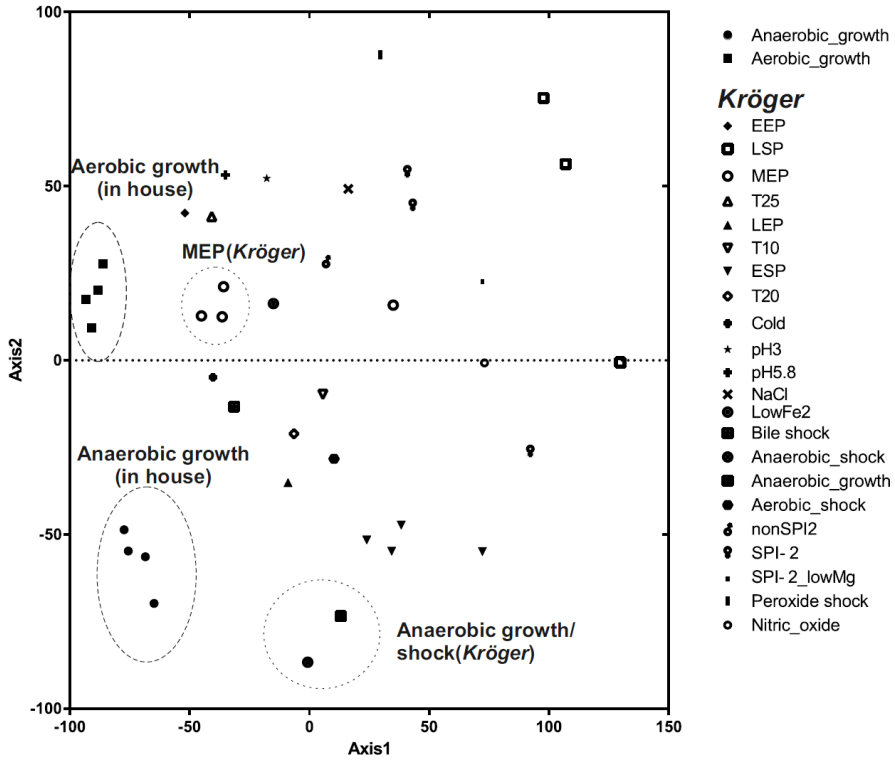
Amongst the *in vivo* samples, *S*. Tm respond to the host gut environment alone in Germ-free conditions and *S*. Tm respond to a combinatorial effect of the host gut environment and microbiota of different compositions in SPF-1 and SPF-2 conditions. In order to identify the *S*. Tm transcriptional systems that were associated with the host gut environment, we compared Germ-free samples to that of the *in vitro* samples and we termed them as host-associated gene signatures of the *S*. Tm. In order to identify the *Salmonella* transcriptional systems that were associated with the microbiota in host gut environment, we compared Germ-free samples to that of the SPF-2 and SPF-1 samples and we termed them as microbiota-associated gene signatures of *Salmonella*.

The genomic features of the *S*. Tm fall into two categories: protein-coding and non-coding RNAs. *S*. Tm transcribe all the genes in a transcriptional unit into one polycistronic mRNA because all the genes within this transcriptional unit perform a function together. If the transcriptional mechanism associated with the genes in a transcriptional unit is assumed active, then all the genes in the transcriptional unit should be active. The non-coding RNAs were largely observed to be performing regulatory activity. Hence, we organized the *S*. Tm transcriptome analysis into two sections: one to identify the potential gene-signatures from the protein coding genes and the other from the non-coding RNAs. The protein coding genes were further organized into the Transcriptional Units (TUs) - gene(s) between one Transcription Start Site and Transcription Stop Site pair, by doing which we reduced the false positive scenarios like those TUs with only one gene showing significant differential expression. We considered only those TUs with consistent expression pattern across all the genes in it to be significant. In addition, here we refer to those TUs with more than one gene as operon.

7.1.1 *S*. Tm transcriptional machinery in aerobic and anaerobic conditions:

Before studying the *modus operandi* of the *S*. Tm infection *in vivo*, we characterized the transcriptional machineries that could be associated with aerobic and anaerobic *in*

vitro conditions. To characterize the transcriptional landscape for these *in vitro* conditions, we compared them with the expression profile of the *S*. Tm obtained under infection-relevant *in vitro* conditions (Kröger et al., 2013). The ordination plot (based on Sammon distance) given below depict this comparison. It shows us that the anaerobic samples from both the experiments had similar expression pattern in comparison to samples from other conditions and the aerobic samples from this study have similar expression pattern to the Mid-Exponential Phase (MEP) from the Kröger et.al., 2013 study, which could be attributed to their similarity in experimental setup. It also shows that there is limited bias based on the technical factors and variation seen between the samples from the two in-house conditions could be attributed to the biological variation.



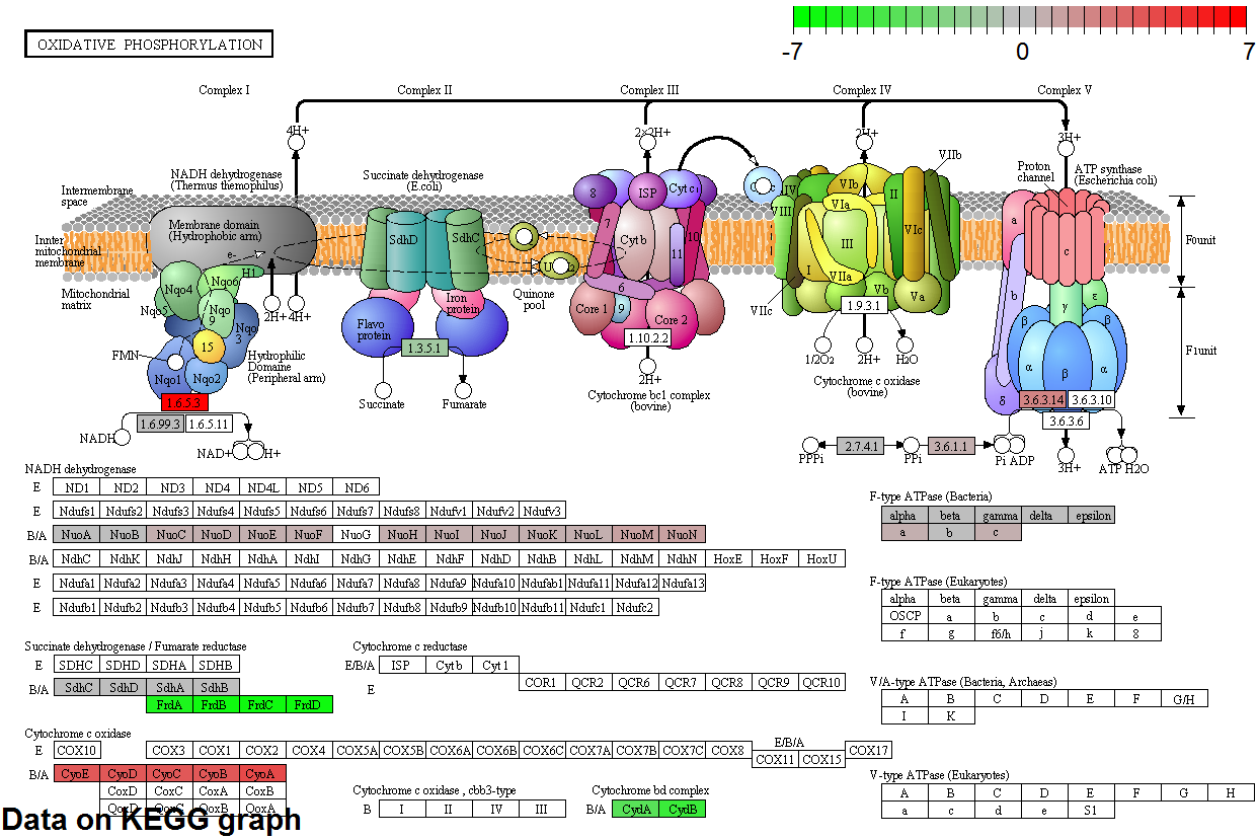
Ordination plot for the *in vitro* samples (from both the in-house data and from Kröger et al., 2013), by normalizing the data based on TU/operon-based TPM normalization and obtaining the Sammon distance between the samples.

Figure 7-2 Sammon plot for the *S*. Tm expression profiles *in vitro*

To characterize the *S*. Tm transcriptome landscape under aerobic and anaerobic *in vitro* conditions, we compared their expression profile and obtained those operons/TUs that were highly expressed (upregulated) in both the conditions. There were 406 TUs (542 genes) highly expressed in aerobic condition and 561 TUs (783 genes) were highly expressed in anaerobic conditions. Below we highlight few interesting patterns observed in this comparison.

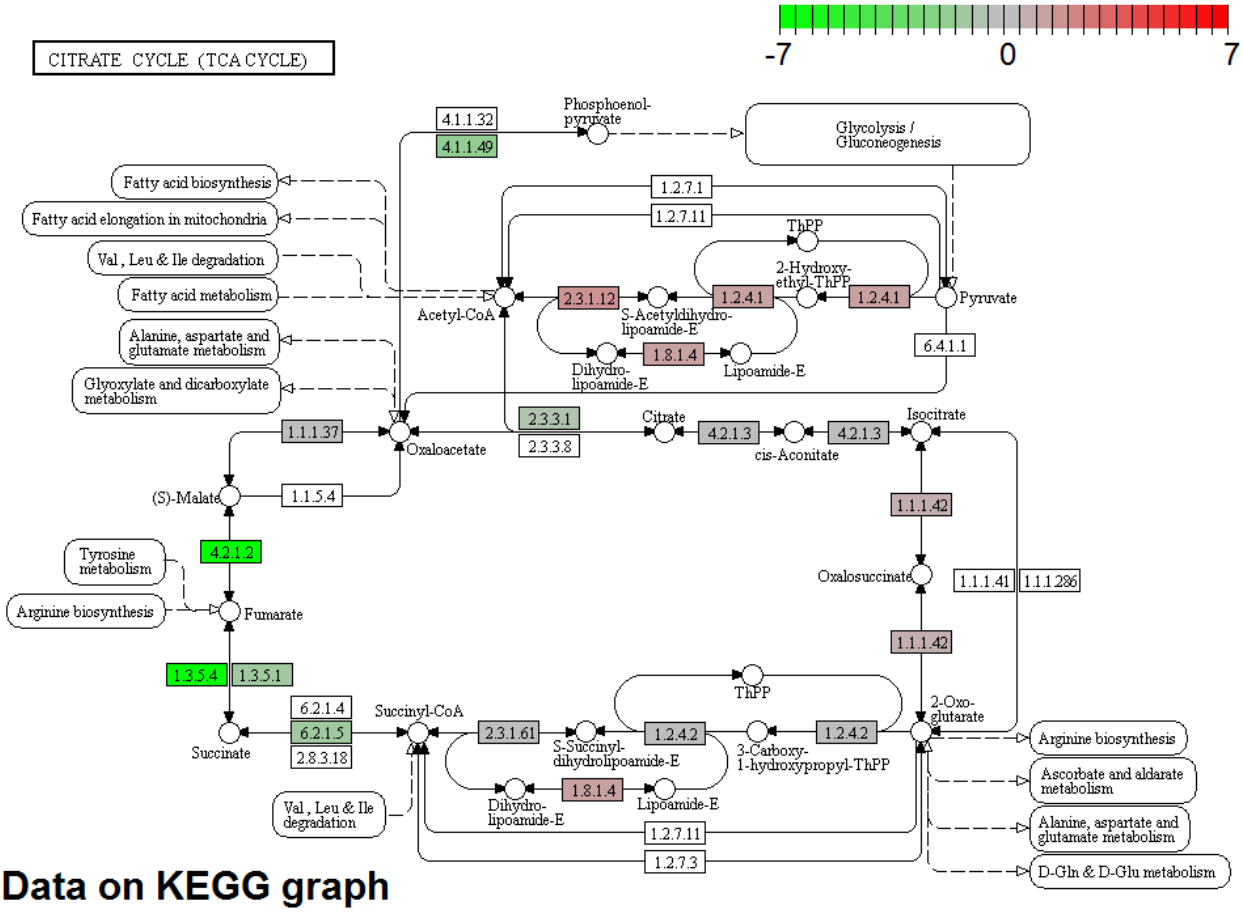
1. **Choice of nutrients**: Under anaerobic condition, *S*. Tm highly expressed operons/TUs responsible for the uptake and utilization of propanediol and ethanolamine, which are known to be the primary carbon source in the host environment (Sinha et al., 2012) for *Salmonella*.
2. **Virulence**: The *S*. Tm upregulated operons responsible for invasion (inv-operon), adhesion to mammalian cells (Nishiyama et al., 2007)-(mannose-dependent type-I fimbriae fim operon) and Type-II Secretion System (pil operon) under anaerobic conditions. Correspondingly, *PhoP-PhoQ* activated virulence proteins *pagK, pagM* were upregulated in aerobic conditions.
3. **Energy metabolism**: The characteristic anaerobic electron transport systems cyd-operon and frd-operon were upregulated in anaerobic condition. Interestingly, *S*. Tm utilized the complete TCA cycle under both aerobic and anaerobic conditions though the succinate dehydrogenase operon (Complex I of theoxidative phosphorylation) was highly expressed in anaerobic condition. *S*. Tm also upregulated cyo Operon involved in electron transport mechanism known to be active in oxygen-rich condition (Iuchi et al., 1990), (Kumar and Shimizu, 2011). These are depicted in the figures **Figure 7-3**, **Figure 7-4**, **Figure 7-5** where the differential expression values were plotted over the KEGG pathways TCA Cycle and Oxidative Phosphorylation.

S. Tm also upregulated operons responsible for nitrate reduction- nar- & nrf operons and the two-component system *narQP* that regulate these operons were upregulated in the anaerobic condition. A previous study has reported operons consisting of *nrdD,nrdG* and *nrdA,nrdB* genes for its requirement in strict anaerobic environment and aerobic environments respectively for ribonucleotide reduction (Garriga et al., 1996). Correspondingly, we observed operons consisting of *nrdD,nrdG* and *nrdA,nrdB* genes upregulated in anaerobic and aerobic environments respectively. In addition, the operons encoding genes for flagellar export apparatus (fli-operon) were upregulated under anaerobic condition.



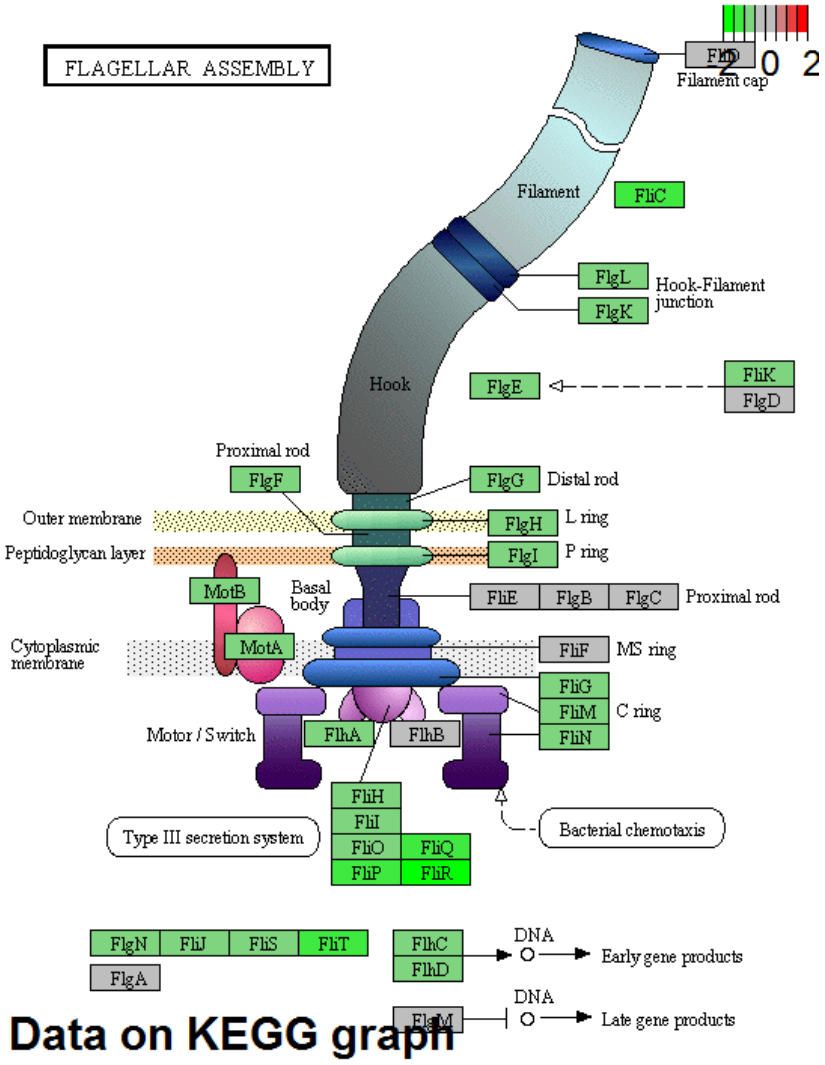
Data on KEGG graph

Figure 7-3 Differentially Expressed Geness of the Oxidative Phosphorylation *in vitro*



Data on KEGG graph

Figure 7-4 Differentially Expressed Genes of the TCA Cycle *in vitro*



We obtained the operons that were differentially expressed by comparing samples grown in aerobic condition against those samples grown in anaerobic condition. We mapped the genes constituting such operons to KEGG Pathway **Figure 7-3** oxidative phosphorylation **Figure 7-4** TCA Cycle **Figure 7-5** Flagellar Assembly.

Figure 7-5 Differentially Expressed Genes of the Flagellar Assembly *in vitro*

It is interesting to note that we observed a similar pattern in expression for the samples obtained from similar experimental set up from (Kröger et al., 2013). That is, irrespective of the difference in studies, the functional systems highly active in the in-house samples grown in aerobic conditions were also highly active in (Kröger et al., 2013) samples grown in mid-exponential phase conditions and the functional systems highly active in in-house samples grown in anaerobic conditions were also highly active in (Kröger et al., 2013) samples grown in anaerobic conditions. This is depicted by the heatmap given below in **Figure 7-6**.

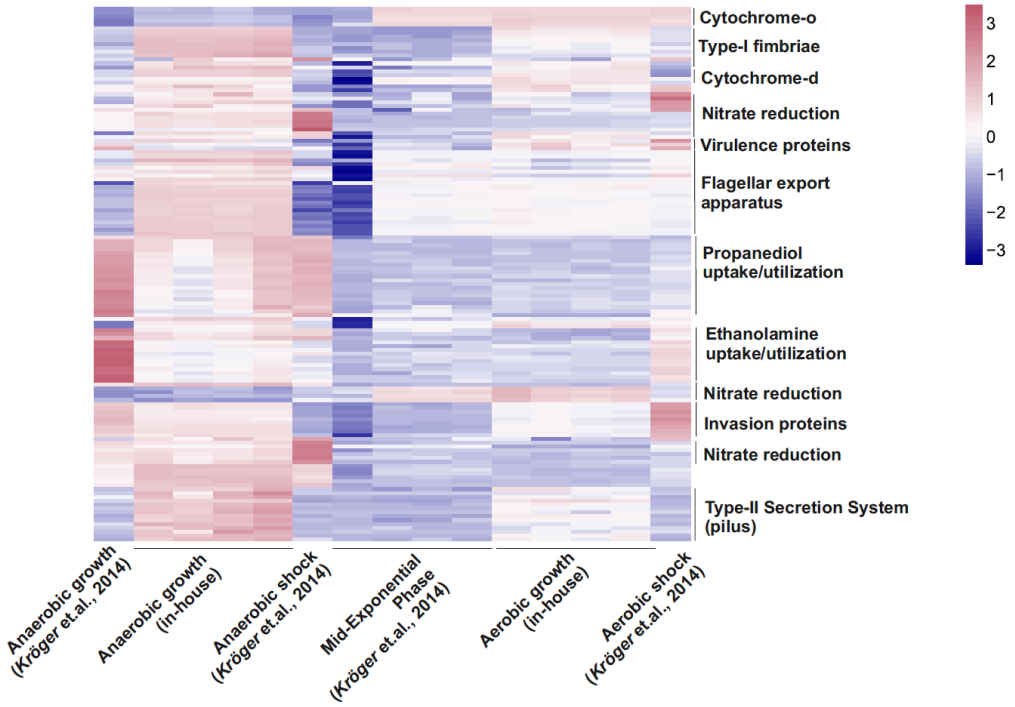


Figure 7-6 Heatmap for Differentially Expressed Genes across all aerobic and anaerobic *in vitro* samples

The genes that were up and down regulated in in-house anaerobic growth samples against in-house aerobic growth samples were taken and the log2-TPM normalized expression profile of these genes were represented as heat map (after normalizing for genes across samples) for samples grown in-house in aerobic and anaerobic conditions and samples grown in mid-exponential phase and aerobic growth conditions from Kröger et.al., 2013 study.

7.1.2 Relevancy of *in vitro* studies with *Salmonella* infection *in vivo*

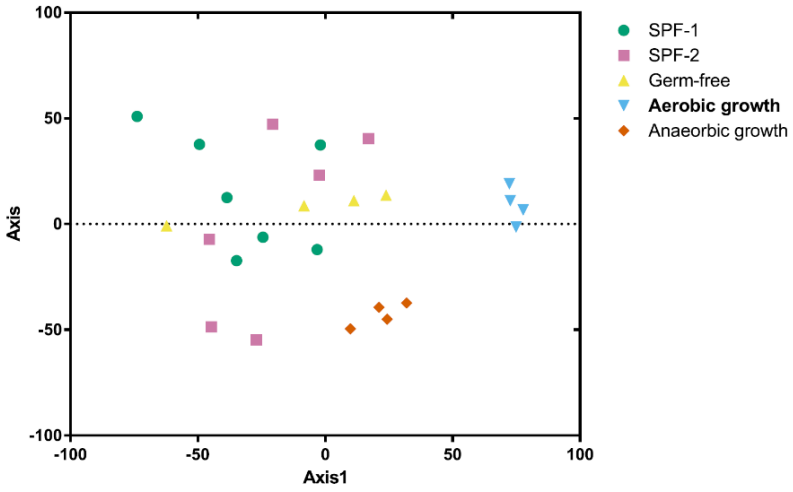


Figure 7-7 PCA plot for the *S*. Tm expression profiles *in vitro* and *in vivo*

Ordination plot for all the in-house samples, by normalizing the data based on TU/operon-based TPM normalization and obtaining the Sammon distance between the samples.

The previous section gave us an indication that *S*. Tm employs distinct functional systems for the experimental *in vitro* conditions- aerobic and anaerobic. Before an in-depth analysis, we compared the expression profile of *S*. Tm *in vivo* to *in vitro* conditions by producing an ordination plot for them as shown above. Based on the ordination plot, one could understand that *S*. Tm *in vivo* transcriptional machinery would be quite distinct when compared to its transcriptional machinery in the *in vitro* conditions. A detailed study on the relevancy of the *in vitro* studies to understand *S*. Tm infection *in vivo* is done in the next section.

7.2 Host-associated signatures

We estimated the number of genes that were active (genes with a median TPM greater than 10) of the *S*. Tm in Germ-Free condition and *in vitro* condition and **Figure 7-8-A** shows a qualitative comparative analysis of the number of genes expressed/not expressed in the *S*. Tm in Germ-free and *in vitro* conditions (irrespective of their scale of expression). We identified 2577 genomic features (i.e. ~50%) were active in all the three conditions and 1676 genomic features (~32%) were not active in all the three conditions. Only 189 (~4%), 63 (~1%) and 155(~3%) genomic features were estimated to be uniquely active in Germ-Free, aerobic and anaerobic growth conditions respectively. This gives an indication that *S*. Tm employ a fraction of its genomic features uniquely for its adaptation against the host defense when compared to that of its growth *in vitro*.

We compared the expression profile of *S*. Tm in *in vitro conditions* (aerobic and anaerobic growth) against its growth in Germ-free *in vivo* condition to identify the potential functional systems that were comparatively active and dormant in the host gut environment. For this, we identified the Transcriptional Units (TUs) that were highly expressed (upregulated) and less expressed (downregulated) in the *S*. Tm expression profile in Germ-free mouse gut environment against both the *in vitro* conditions (Anaerobic and Aerobic). Furthermore, we looked for the functional and sub-functional categories they enrich.

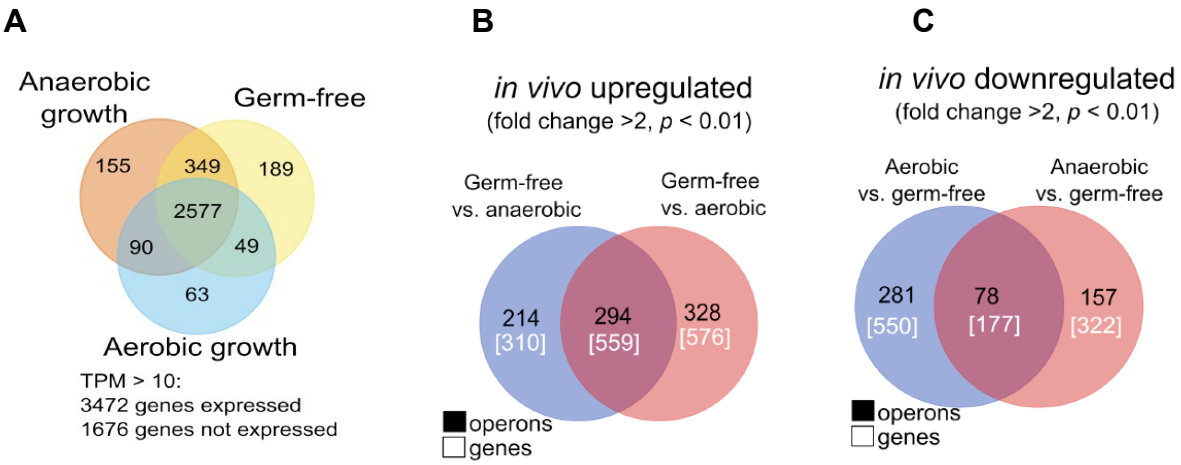


Figure 7-8 Comparative analysis of *S.* Tm expression in Germ-Free and *in vitro*

A. The venn diagram compare the estimated number of genes expressed in Germ-free, Aerobic and Anaerobic *in vitro* conditions to each other. For this, we obtained the average TPM normalized expression of a gene in each condition. We estimated the number of genes expressed by the assumption that a gene with TPM>10 to be expressed in FANTOM consortium. B. We organized genes into the transcriptional units they belong and those TUs with average fold change >=2 or Kost's p-value <1 % (see Materials and Methods for detailed description) were considered to be differentially expressed. We compared the TUs expressed higher in samples from aerobic condition than those samples in Germ-free condition and that of anaerobic against Germ-free conditions and represented them as venn diagram. The numbers in black indicate the upregulated TUs and the number in white indicate the upregulated genes. The TUs at the intersection of the venn diagram represent the potential *in vitro*-associated operons/genes. After obtaining the venn diagram, we removed the operons that were subset of another operon in the same venn group. C. Similar to B, we obtained the TUs expressed higher in Germ-free samples than those in aerobic samples and that of Germ-free samples against anaerobic samples and represented them as venn diagram. The numbers in black and white represent upregulated operons and genes respectively. The TUs at the intersection of the venn diagram represent the potential Germ-free-associated operons/genes. After obtaining the venn diagram, we removed the operons that were subset of another operon in the same venn group.

conditions

First, we identified those *S.* Tm TUs that were upregulated in Germ-free condition when compared to both the *in vitro* conditions. The venn diagram in **Figure 7-8-B** depict this comparison. There were 294 non-overlapping, unique upregulated TUs in *S.* Tm Germ-free condition (Fold change ≥ 2). These 294 TUs correspond to 559 protein-coding genes. We have compiled the interesting expression pattern from these genes.

Pathogenesis machinery:

1. **SPI-1 genes upregulation**: We observed 25 genes that were involved in SPI-1 T3SS machinery. The SPI-1 T3SS is characterized on forming syringe-like organelles on the *S.* Tm surface and inject proteins directly into the eukaryotic cells (Dieye et al., 2009) (supported by upregulation of the master regulator hilD).
2. **Colicin-plasmid genes upregulation**:
    a. There were 30 unique TUs in the virulence plasmid, out of which six were active in Germ-Free condition compared to the *in vitro* conditions. This

corresponds to 40 genes. These genes are associated with the following functions:

b. Icm/Dot secretion system proteins, a type of Type-IV Secretion System (tra-operon) which are the conjugation machineries that transfers macromolecules via direct cell-to-cell contact (Christie et al., 2014) were upregulated in the Germ-free condition.

c. Pilus proteins and the associated shufflon protein were upregulated in Germ-free condition compared to the *in vitro* conditions (Morris et al., 2003).

d. Type-1 Toxin Antitoxin proteins: *pndA,pndC* genes present in the colicin plasmid was upregulated in the Germ-free condition.

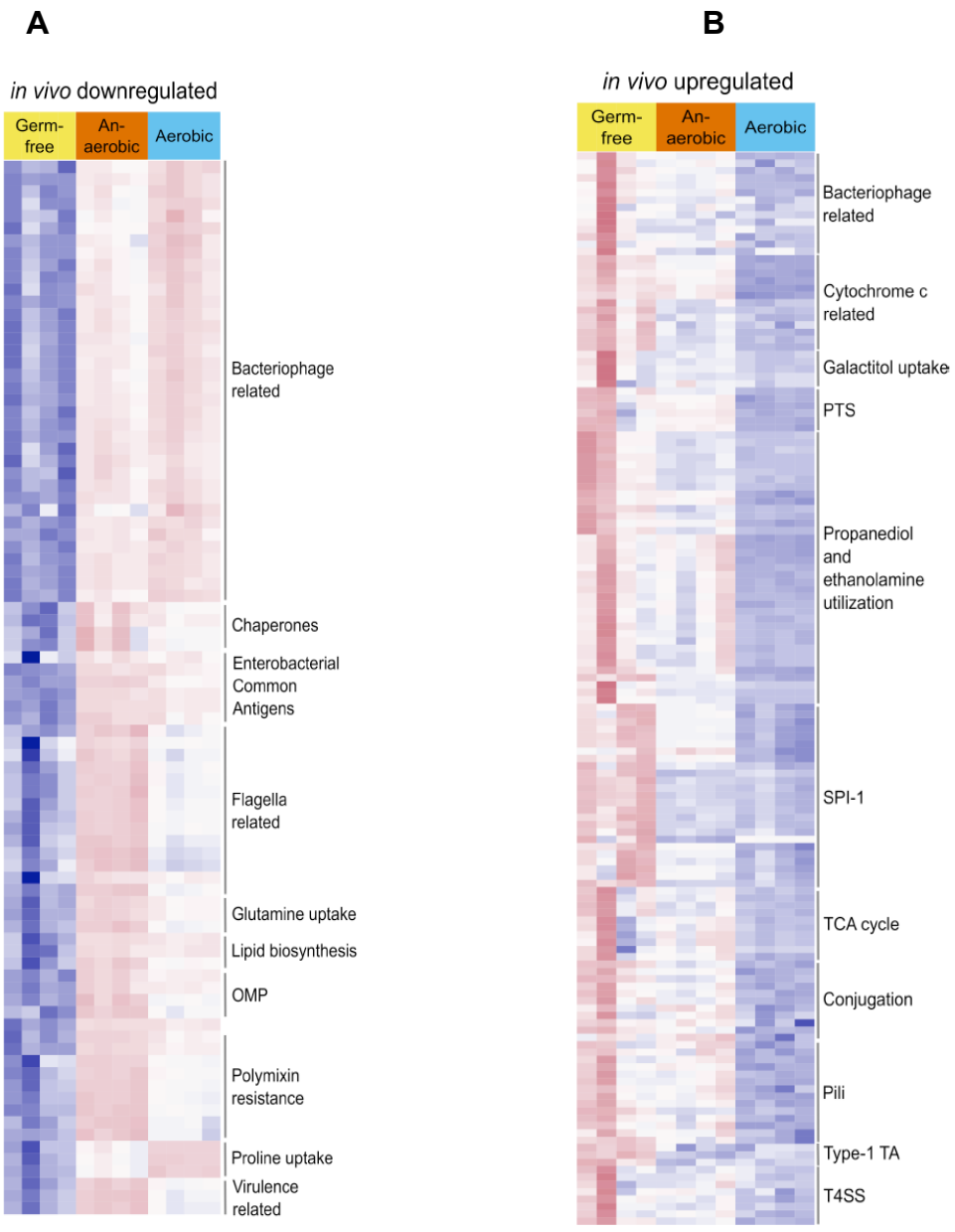


Figure 7-9 Heatmap of the host-associated *S*. Tm gene signatures

The genes that were up and down regulated in Germ-free samples against samples from both the aerobic and anaerobic in vitro conditions (represented by the intersection part Figure 8-8B and Figure 8-8C respectively) were taken and the log2-TPM normalized expression profile of these genes were represented as heat map (after normalizing for genes across samples). *A.* We handpicked those operons that were downregulated in Germ-.free samples against samples from both the aerobic and anaerobic conditions and (represented in the intersection of the venn diagram Figure 8-8B) represented it as heatmap. Similarly, we handpicked those operons that were upregulated in Germ-.free samples against samples from both the aerobic and anaerobic conditions (represented in the intersection of the venn diagram Figure 8-8C) and represented it as heatmap.

3. **Host-associated metabolic niches**:
    a. The propanediol and ethanolamine utilization upregulation (master regulator *pocR,eutK, eutR*) , (Sinha et al., 2012), (Srikumar and Fuchs, 2011) : It has been reported before that *S*. Tm is capable of utilizing 1,2-propanediol and ethanolamine as the sole carbon source in host gut environment (Srikumar and Fuchs, 2011). Given that the *S*. Tm had no competition for the propanediol and ethanolamine in the Germ-free gut environment, our observation on the upregulation of the uptake and utilization systems of the 1,2-propanediol and ethanolamine in comparison to the *in vitro* conditions further evaluates to this argument. Ethanolamine, which is known to be produced by the breakdown of the mammalian cell membranes (Srikumar and Fuchs, 2011), could be utilized as a sole energy source in *S*. Tm. The tetrathionate (electron acceptors for 1, 2-propanediol) and the master regulator *hilD* which were associated with the propanediol and ethanolamine utilization was also upregulated.
    b. L-ascorbate metabolism: The operons that were associated with the L-ascorbate (a glucose derivative) uptake and utilization (sga- operon,,yia-operon, lyx-operon and sgb-operon) (Yew and Gerlt, 2002), (Campos et al., 2008), (Ibañez et al., 2000) were upregulated in the Germ-free condition.
4. **Energy metabolism(Electron transport)**:
    a. The operons consisting of cytochrome-c type proteins (nap-operon,nrf-operon and ccm-operon) were upregulated in the Germ-free condition. These operons were previously considered to be expressed under aerobic conditions. This operon is constituted by genes *napA, napB* that are responsible for receiving electrons from *napC*, allowing electron flow between membrane and periplasm. Ten out of the 26 genes involved in TCA Cycle in the *S*. Tm were upregulated in Germ-Free condition. These

genes mediate the conversion of 2-oxoglutarate to malate. This includes the sdh-operon, suc-operon (Complex II of Oxidative Phosphorylation), and fum-operon. Previous studies have showed that deletion of these operons could be associated with the avirulence of the *S*. Tm (Mercado-Lubo et al., 2008).

Similarly, we identified the TUs that were downregulated in the Germ-free condition against the *in vitro* conditions, when compared separately. This scenario is shown in **Figure 7-8-C**. There were 78 non-overlapping and unique TUs (177 genes) downregulated in the *S*. Tm Germ-free expression profile against the *in vitro* profiles. We highlight the interesting gene expression patterns from these genes below.

1. Cationic antimicrobial peptides resistance upregulation: The *pmrF* Operon (with seven genes) and *BasS-BasR* (Barchiesi et al., 2009) operon, was downregulated in the Germ-Free condition when compared to the *in vitro* conditions.
2. Downregulation of the flagellar genes (downregulation of the regulator *fliZ*, which could be associated with the upregulation of hilD (Singer et al., 2014)) Previously, it is reported to play a role in *S*. Tm infection in pigs, but not in chicken or calf ((Chaudhuri et al., 2013)). Here, we observed the flagellar system to be less expressed in the Germ-free condition. The rff-operon that is responsible for enterobacterial common antigen biosynthesis and thereby protection against bile salts were downregulated in Germ-Free condition. These polysaccharides were also previously studied to be associated with the flagellar motility (Gilbreath et al., 2012).

Thus, *S*. Tm responds to the Germ-free host gut environment by upegulating transcriptional machineries associated with the host-specific nutrients uptake/utilization, virulence and conjugal transfer proteins. However, it downregulated (not switched off) the flagellar operons associated with motility and lipopolysachcharide biosynthesis to evade the host immune defense. This section also emphasis the observation that such *in vitro* studies could only give us an indication on the transcriptional machineries association with *S*. Tm virulence *in vivo*.

7.3 Microbiota-associated signatures

In this section, we studied how *S*. Tm respond to different gut microbiota compositions. For this, we compared the expression profile of *S*. Tm in microbiota-associated *in vivo* conditions (SPF-1 and SPF-2 gut microbial communities) against its growth in the Germ-free mouse gut environment to identify the potential functional systems that were comparatively active and dormant in the competitive host environment.

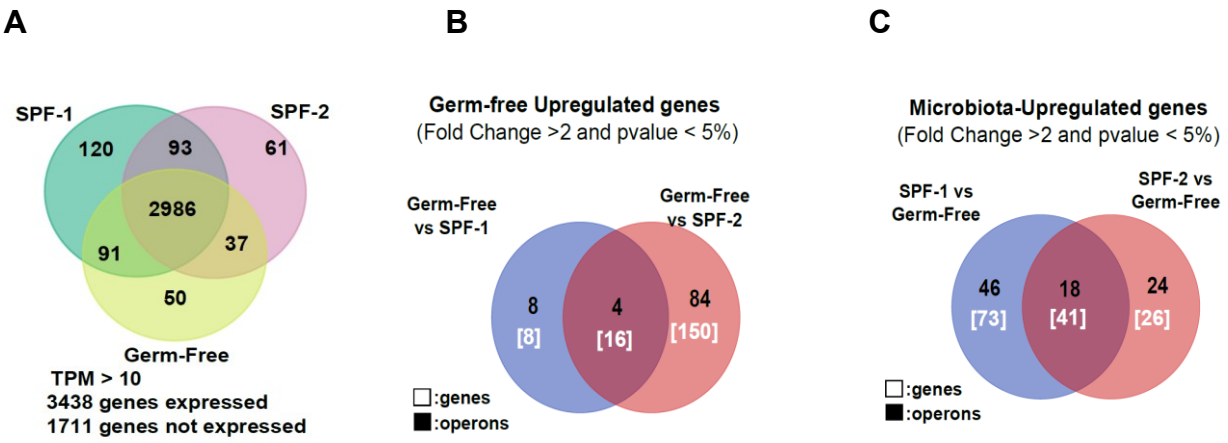


Figure 7-10 Comparative analysis of *S*. Tm expression *in vivo*

A. The venn diagram compare the estimated number of genes expressed in Germ-free, SPF-1 and SPF-2 *in vivo* conditions to each other. For this, we obtained the average TPM normalized expression of a gene in each condition. We estimated the number of genes expressed by the assumption that a gene with TPM>10 to be expressed in FANTOM consortium. B. We organized genes into the transcriptional units they belong and those TUs with average fold change >=2 or Kost's p-value <1 % (see Materials and Methods for detailed description) were considered to be differentially expressed. We compared the TUs expressed higher in samples from SPF-1 condition than those samples in Germ-free condition and that of SPF-2 against Germ-free conditions and represented them as venn diagram. The numbers in black indicate the upregulated TUs and the number in white indicate the upregulated genes. The TUs at the intersection of the venn diagram represent the potential *in vivo*-associated operons/genes. After obtaining the venn diagram, we removed the operons that were subset of another operon in the same venn group. C. Similar to B, we obtained the TUs expressed higher in Germ-free samples than those in SPF-1 samples and that of Germ-free samples against SPF-2 samples and represented them as venn diagram. The numbers in black and white represent upregulated operons and genes respectively. The TUs at the intersection of the venn diagram represent the potential Germ-free-associated operons/genes. After obtaining the venn diagram, we removed the operons that were subset of another operon in the same venn group.

Figure 7-10-A shows the qualitative comparative analysis of the number of genes expressed and not expressed in the *S*. Tm expression profile of all the *in vivo* conditions (irrespective of their scale of expression). 2986 genomic features (~58%) of the 5149 genomic features were expressed in all the conditions. Only 50 (~1%), 120 (~2%) and 61 (~1%) of the total genomic features were uniquely expressed in Germ-free, SPF-1 and SPF-2 *in vivo* conditions.

Similar to the method adapted in identifying host-associated signatures, we compared the upregulated and downregulated TUs in Germ-free when compared to SPF-1 and SPF-2 conditions. **Figure 7-10-B** and **C** shows the qualitative comparative analysis of the number of TUs and the corresponding genes upregulated in the pair-wise comparisons. There were 4 non-overlapping and unique TUs (16 genes) that were upregulated in Germ-free condition.

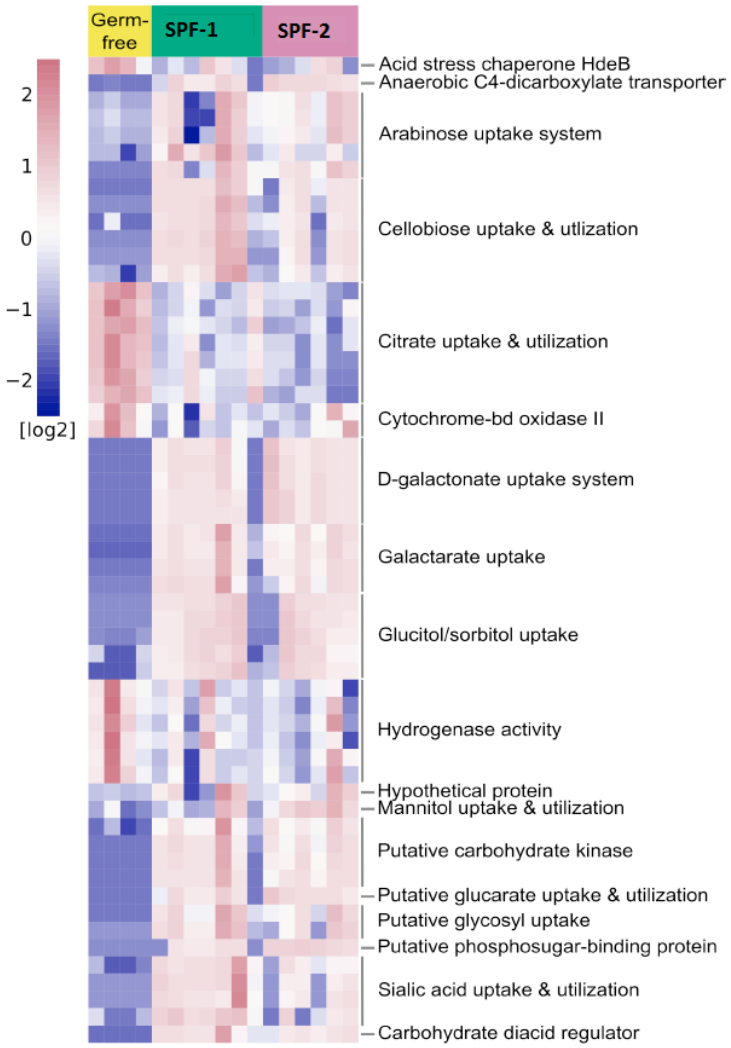


Figure 7-11 Heatmap of the microbiota-associated *S*. Tm gene signatures

The operons/TUs that were up and down-regulated in Germ-free samples against samples from both the SPF-1 and SPF-2 *in vivo* conditions (represented by the intersection part of Figure 8-10B and Figure 8-10C respectively) were taken and the log2-TPM normalized expression profile of these genes were represented as heat map(after normalizing for genes across samples). We obtained those operons that were up and downregulated in Germ-.free samples against samples from both the SPF-1 and SPF-2 conditions and represented it as heatmap.

Microbiota-associated signatures downregulated in *S.* Tm:

We observed the citrate uptake/utilization system, hydrogenase activity, cytochrome- bd oxidase and acid stress response associated transcriptional units were upregulated in the Germ-free condition compared to the SPF-1 and SPF-2 gut microbiota conditions. Acid stress response protein hdeB, which was upregulated in the Germ-free condition compared to the colonized mice samples, is known to counteract the acidic nature of the digest tract (Zhao and Houry, 2010). NiFe oxygen tolerant Hydrogenase operon (hya-operon) that was upregulated in the Germ-free condition when compared to both the colonized mice conditions, was previously studied to be associated with the virulence of *S*. Tm *in vivo* (Maier, 2005) .and required under anaerobic conditions. In addition, the cytochrome-bd oxidase known for protection against the hydrogen peroxide and nitric oxide stress was also upregulated in the Germ-free conditions. The citrate uptake/utilization operon, (cit-operon) which uptake and degrade citrate to oxaloacetate were upregulated in the Germ-free condition.

Microbiota-associated gene signatures upregulated in *Salmonella*:

Figure 7-10-C shows the qualitative comparative analysis of the number of TUs and the corresponding genes downregulated in the pair-wise comparisons. There were 18 non-overlapping and unique TUs (41 genes) that were upregulated in the *in vivo* conditions compared to the Germ-free conditions. We highlight the interesting expression pattern from these genes below.

1. When compared to the Germ-free conditions, there were quite a number of Transcriptional Units (TUs) associated with the carbohydrate uptake/utilization systems upregulated in the *S*. Tm in the colonized mice conditions. This could be attributed to the ability of *S*. Tm to utilize the available nutrients for energy production and metabolism. Predominantly, most of them were simple sugars (cellobiose, glucarate, sorbitol, galactarate, glucitol,glycosyl and arabinose) and also hypothetical carbohydrate uptake/utilization systems. This could indicate that in the colonized mouse models, due to the competition from other gut microbiota, *S*. Tm face a carbon-limited environment. Carbohydrate diacid regulator protein *cdaR* and the nutrient-acquisition systems glucarate, galactarate that were proposed to be regulated by it in a study in *E*. coli (Monterrubio et al., 2000), were also upregulated in the colonized mice conditions when compared to the Germ-free conditions. A previous study has also proposed an association between Hydrogen-fuel and glucarate uptake/utilization system and glyoxylate pathway (Lamichhane-Khadka et al.,

2011) which were upregulated in *S*. Tm in the colonized mice conditions in our in-house data.

From the identified *in vivo* protein-coding gene signatures, we could infer that *S*. Tm adaptation is largely dependent on the microenvironments in the gut environment, which is resultant of the combinatorial effect of the host and the SPF-1 and SPF-2 gut microbiota.

7.4 Non-coding genes

From the transcriptome analysis of the protein-coding regions, one could infer that the *S*. Tm transcriptional machineries associated with the nutrient uptake/utilization and associated specific regulator, energy metabolism and virulent systems were differentially expressed *in vivo*. In this section, we identified the differentially expressed *S*. Tm non-coding RNAs and the possible associations to the *S*. Tm protein-coding gene signatures identified in the previous sections. Most of the *S*. Tm non-coding RNAs were associated with the transcriptional regulatory activity. We analyzed the expression profile of the *S*. Tm non-coding RNAs (280 non-coding RNAs mentioned in (Kröger et al., 2013) and the known and predicted non-coding RNAs from (Kröger et al., 2012)). First, we checked whether the non-coding RNA length were independent of the number of reads mapped by plotting the length against the number of reads mapped to all the samples. The **Figure 7-12**-**A** showed that the reads mapped to the non-coding RNAs were independent of the length of non-coding RNAs i.e. the reads mapped to the non-coding RNAs were unbiased for the length of them thereby confirming that size selection performed during RNA-seq library preparation had little or no effect on quantifying the non-coding RNAs of small size. **Figure 7-12**-**B** is a bar graph of the fraction of the transcripts belonging to non-coding RNAs (TPM normalized) for each condition. In *in vitro* conditions, around 11% and 13% of the TPM values were from the *S*. Tm non-coding RNAs of aerobic and anaerobic conditions respectively. In the *in vivo* conditions, around 21%, 24% and 20% of the overall TPM values of the samples were from the *S*. Tm non-coding RNAs of GF, SPF-1 and SPF-2 conditions respectively. This shows that the *S*. Tm non-coding RNAs express more *in vivo* compared to that of *in vitro* and they play a significant role *in vivo*. First, we pooled all the *in vitro* samples together and *in vivo* samples together and performed t-test. We obtained the fold change and corresponding p-values for each non-coding RNA between the *in vitro* and *in vivo* conditions. **Figure 7-13** depicts this comparison. Since we wanted to better categorize the differentially expressed non-coding RNAs, we

assigned negative sign to those that were upregulated *in vitro*. To further enrich this list, we compared their fold change against their expression value from the condition where it is comparatively highly expressed. In contrast to the MA plot where they plot log2fold change vs mean abundance, this comparison give us an easy way to identify the *S*. Tm non-coding RNAs that were unexpressed in one condition and highly expressed in another condition. We hereby compile few known and putative non-coding RNAs from this calculation that showed significant difference in expression (p-value≤5% and fold change≥4).

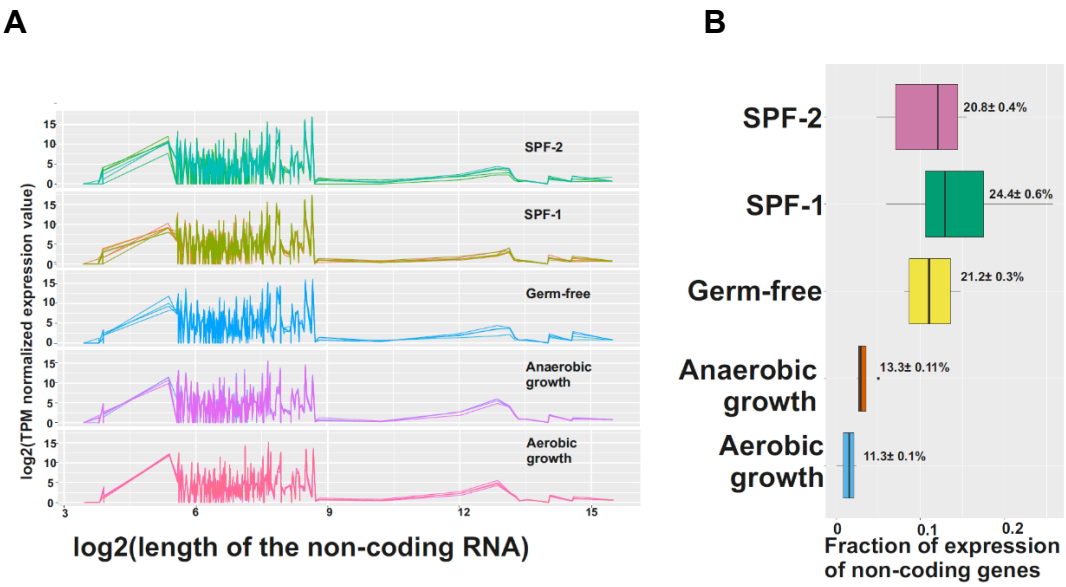


Figure 7-12 Expression levels of *S*. Tm non-coding RNAs across all the conditions under study

A. We sorted the length of the non-coding RNA incrementally (x-axis) and plotted them against their corresponding expression values (y-axis) for each condition. The samples were split into the conditions they belong. *B*. The boxplot represents the fraction of reads (in TPM) that belong to non-coding RNAs. The numbers written denote the mean percentage (with one std.deviation) of reads belonging to non-coding RNAs.

Carbon metabolism: We observed *S*. Tm upregulate multiple nutrient uptake systems under different *in vivo* conditions, possibly be due to competition for the highly preferred simple sugars in the gut environment. As a result, there would be a build-up of sugar phosphate stress within *Salmonella*. We identified few evidences on how the *S*. Tm tackle this issue. We observed SgrS, CsrB known to be involved, and modulating carbon metabolism were upregulated *in vivo* compared against *in vitro*.

1. SgrS: One of the Hfq-associated non-coding RNAs-SgrS, which is believed to be relatively active under high sugar-phosphate conditions (Vanderpool and Gottesman, 2004) in nutrient deprived conditions, was found to be

comparatively upregulated (~40 folds) in all the *in vivo* conditions against *in vitro* conditions.

2. SsrS: It is known to regulate RNA polymerase. It has been shown to be associated with long term survival under nutrient deprived conditions in *E.*coli (Trotochaud and Wassarman, 2004). This non-coding RNA expressed higher *in vivo* (~14 folds) in comparison to the *in vitro* conditions.
3. Stress response: OxyS is a small RNA whose overexpression is associated with oxidative stress response (Altuvia et al., 1997) and with decrease in motility (De Lay and Gottesman, 2012). We also observed a downregulation of the flagellar operons in the Germ-free condition in comparison to *in vitro* conditions. RybA is another small RNA that was overexpressed *in vivo* (~10 folds) in comparison to the *in vitro* conditions. It is studied to be associated with oxidative and/or the peroxide stress (Gerstle et al., 2012).
4. CsrB and CsrC: The expression levels of CsrB and CsrC was very high *in vivo* conditions when compared to *in vitro* conditions, whereas the regulatory gene CsrA was low in both the *in vivo* and *in vitro* conditions. Previous studies have already described about the over-expression of CsrB *in vivo* and potentially act as an antagonist of CsrA (Liu and Romeo, 1997, Liu et al., 1997).It is also previously reported that CsrB overexpression in *E.*coli could be associated with higher intracellular glycogen levels (Liu et al., 1997). It is also interesting to note that, another non-coding RNA RygD which is predicted to interact with CsrA (Sridhar et al., 2009) was also overexpressed *in vivo* conditions in comparison to the *in vitro* conditions.

The non-coding RNAs that were comparatively upregulated *in vitro* conditions were known to be active under nutrient-rich conditions.

5. t44: This non-coding RNA was active *in vitro* conditions which is known to be expressed in actively growing environment (Ortega et al., 2012).

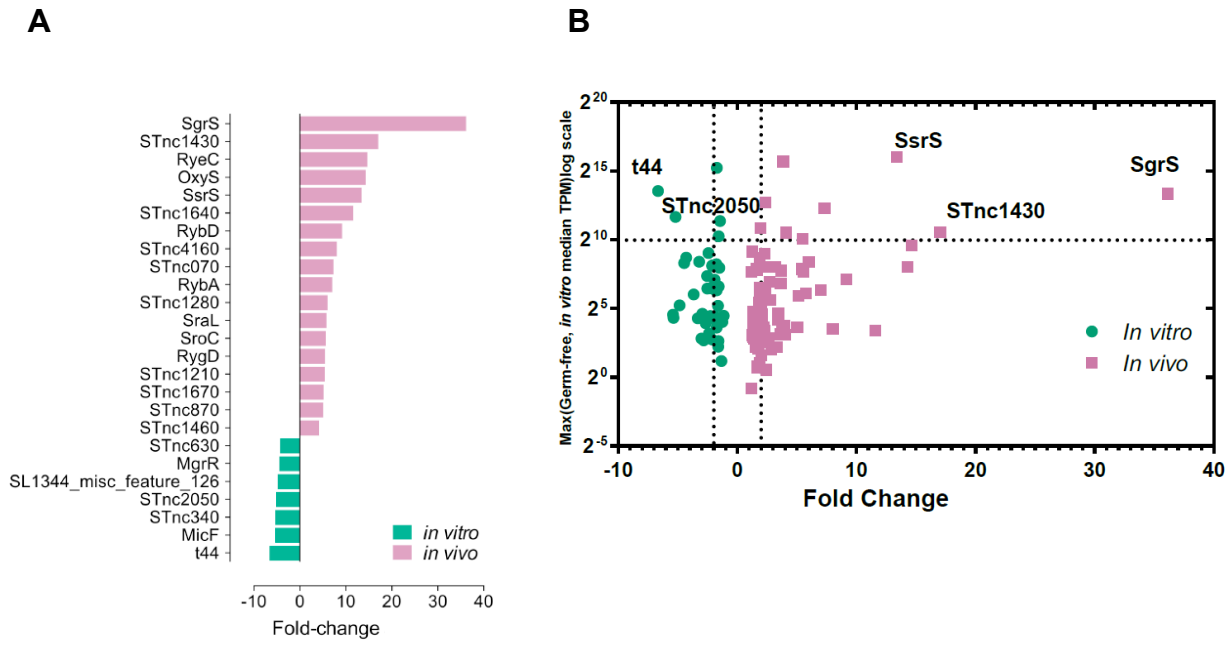


Figure 7-13 non-coding RNAs essential for *S*. Tm adaptation *in vitro* and *in vivo*

A. We compared the expression profile of all the *in vitro* samples against *in vivo* samples and picked only those with a fold change of greater than 3 in either *in vitro* or *in vivo* conditions and p-value <5% . X-axis denotes fold change where the positive fold change denote *in vivo* upregulated and the negative fold change denote *in vitro* upregulated and y axis denote the non-coding RNA. B. We obtained the fold change between *in vivo* and *in vitro* samples and we enriched only those with a p-value <1%. We plotted the fold change along x-axis and the maximum of the *in vivo* and *in vitro* median TPM values for each non-coding RNA along y-axis. We assigned negative signs to those non-coding RNAs, which were overexpressed *in vitro*. A line is drawn for y=±2 to distinguish those with fold change greater and lesser than 2. A line is drawn for x=1000 to identify those genes with a median TPM value ≥ 1000 in either *in vivo* or *in vitro* conditions.

Similarly, we did similar analysis by comparing the *S*. Tm expression profile of samples from Germ-free conditions against *in vitro* conditions. We obtained many small RNAs with unknown function to be overexpressed in the Germ-free condition and few with known functions like SgrS, SsrS and t44 as depicted in **Figure 7-14-A,B**. When we did similar analysis for the comparison of *S*. Tm expression profile between SPF-1 and SPF-2 conditions, we obtained few non-coding RNAs overexpressed in SPF-1 condition, but we did not obtain any non-coding RNAs overexpressed in SPF-2 condition.

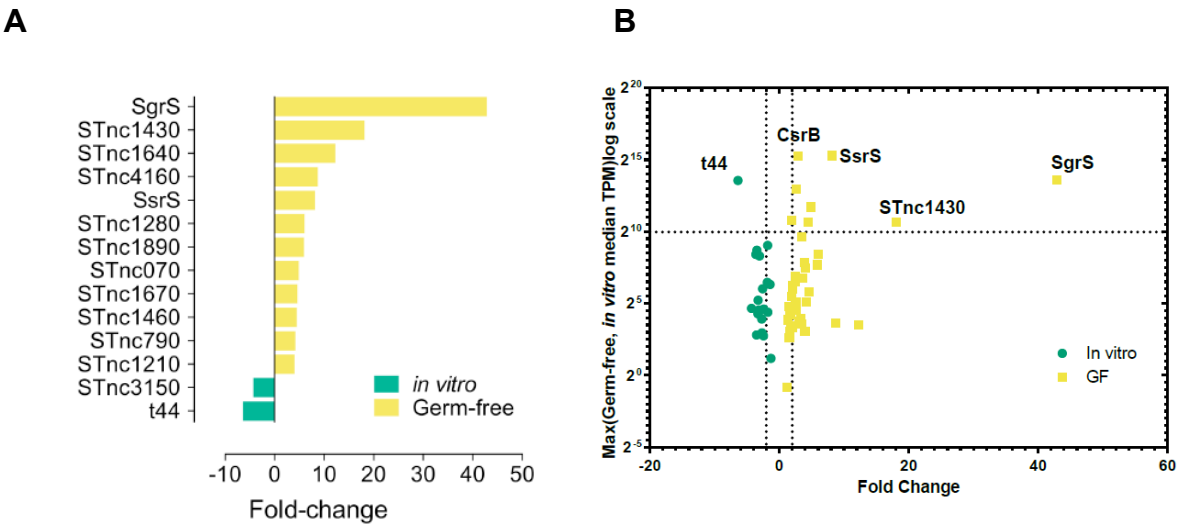


Figure 7-14 non-coding RNAs essential for *S*. Tm adaptation *in vitro* and Germ-Free conditions

A. We compared the expression profile of all the *in vitro* samples against Germ-free samples and picked only those with a fold change of greater than 3 in either *in vitro* or Germ-free conditions and p-value <5% . X-axis denotes fold change where the positive fold change denote Germ-free upregulated and the negative fold change denote *in vitro* upregulated and y axis denote the non-coding RNA. *B*. We obtained the fold change between Germ-free and *in vitro* samples and we enriched only those with a p-value <1%. We plotted the fold change along x-axis and the maximum of the Germ-free and *in vitro* median TPM values for each non-coding RNA along y-axis. We assigned negative signs to those non-coding RNAs, which were overexpressed *in vitro*. A line is drawn for y=±2 to distinguish those with fold change greater and lesser than 2. A line is drawn for x=1000 to identify those genes with a median TPM value ≥ 1000 in either Germ-free or *in vitro* conditions.

A similar comparison was study for Germ-free samples against those from colonized mouse models (SPF-1 and SPF-2 conditions). Many non-coding RNAs with unknown functions like STnc930 and STnc870 were highly expressed (Fold Change >3 & p-value<5%) in the colonized mouse models in comparison to Germ-free condition. However, their expression levels were very low in the colonized mouse models (as shown in y-axis of the figure **Figure 7-15** shown below). CsrC with a fold change slightly higher than 2 was highly expressed in the Germ-free condition (~7000 reads). Also, InvR and STnc2030 small RNAs which had a high fold change value (>5) were having less expression in both the germ-free and colonized mouse model samples. Hence, this might not be a gene signature that could be associated with *S*. Tm virulence in the colonized mouse models. Thus, our way of comparative study allow us to remove potential false positives in identification of *S*. Tm gene signatures.

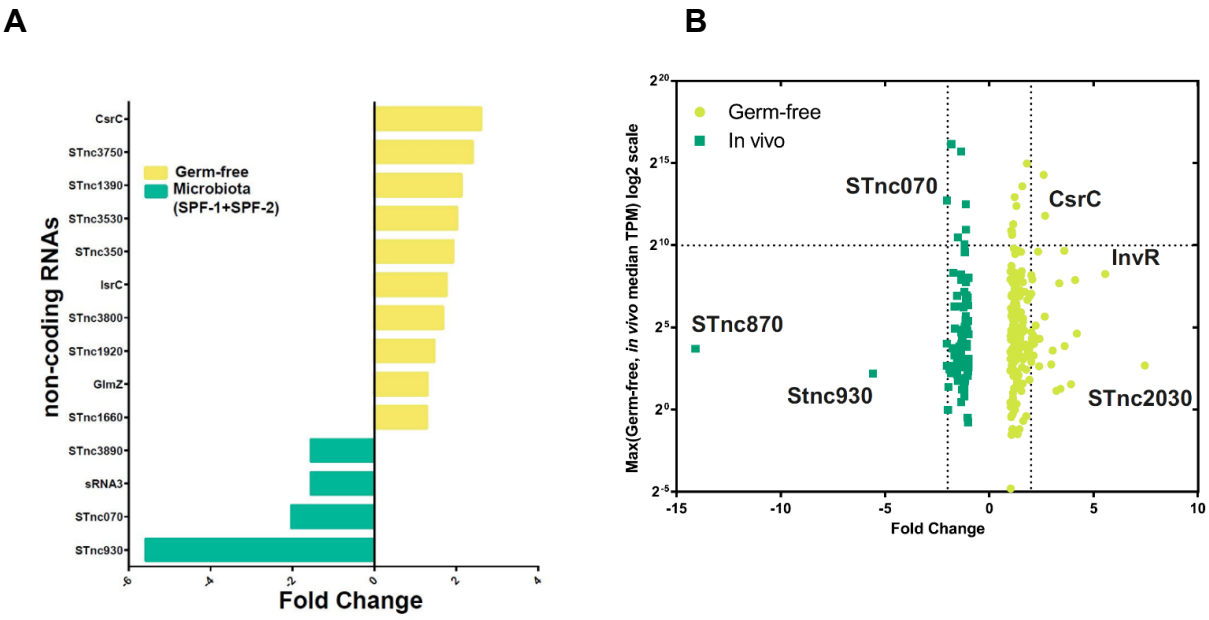


Figure 7-15 non-coding RNAs essential for *S*. Tm adaptation in colonized mouse models and Germ-Free conditions

A. We compared the expression profile of all the *in vivo* samples against Germ-free samples and picked only those with a fold change of greater than 3 in either *in vivo* or Germ-free conditions and p-value <5% . X-axis denotes fold change where the positive fold change denote Germ-free upregulated and the negative fold change denote *in vivo* upregulated and y axis denote the non-coding RNA. B. We obtained the fold change between Germ-free and *in vivo* samples and we enriched only those with a p-value <1%. We plotted the fold change along x-axis and the maximum of the Germ-free and *in vitro* median TPM values for each non-coding RNA along y-axis. We assigned negative signs to those non-coding RNAs, which were overexpressed *in vitro*. A line is drawn for y=±2 to distinguish those with fold change greater and lesser than 2. A line is drawn for x=1000 to identify those genes with a median TPM value ≥ 1000 in either Germ-free or *in vitro* conditions.

Based on the interesting gene signatures we obtained from the protein-coding and non-coding RNA analysis we performed in the previous sections, we compiled them to obtain the potential differential transcriptional regulatory mechanisms employed by the S. Tm under the Germ-free condition and colonized mouse models. In colonized mouse models, we infer that the *S*. Tm is ready to feed on multi-nutrient sources at once for energy metabolism. This could be attributed to the competition provided by other microbes in obtaining those nutrient sources in the gut environment. One interesting thing that could be inferred from the differential transcriptional regulation network is the upregulation of the genes involved in the conversion of all the nutrient sources into a simple sugar-phosphate, which could be further used for central carbon metabolism.

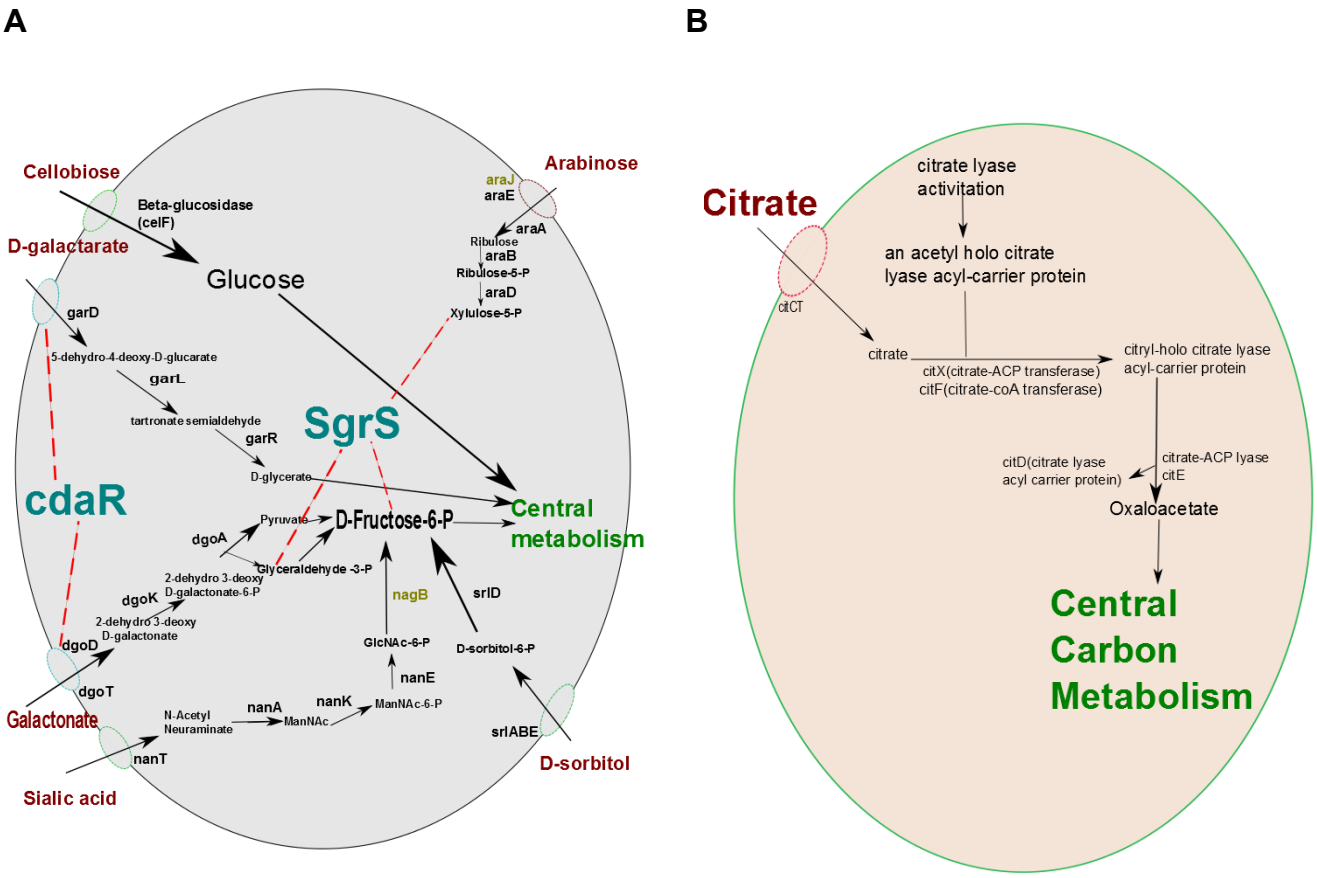


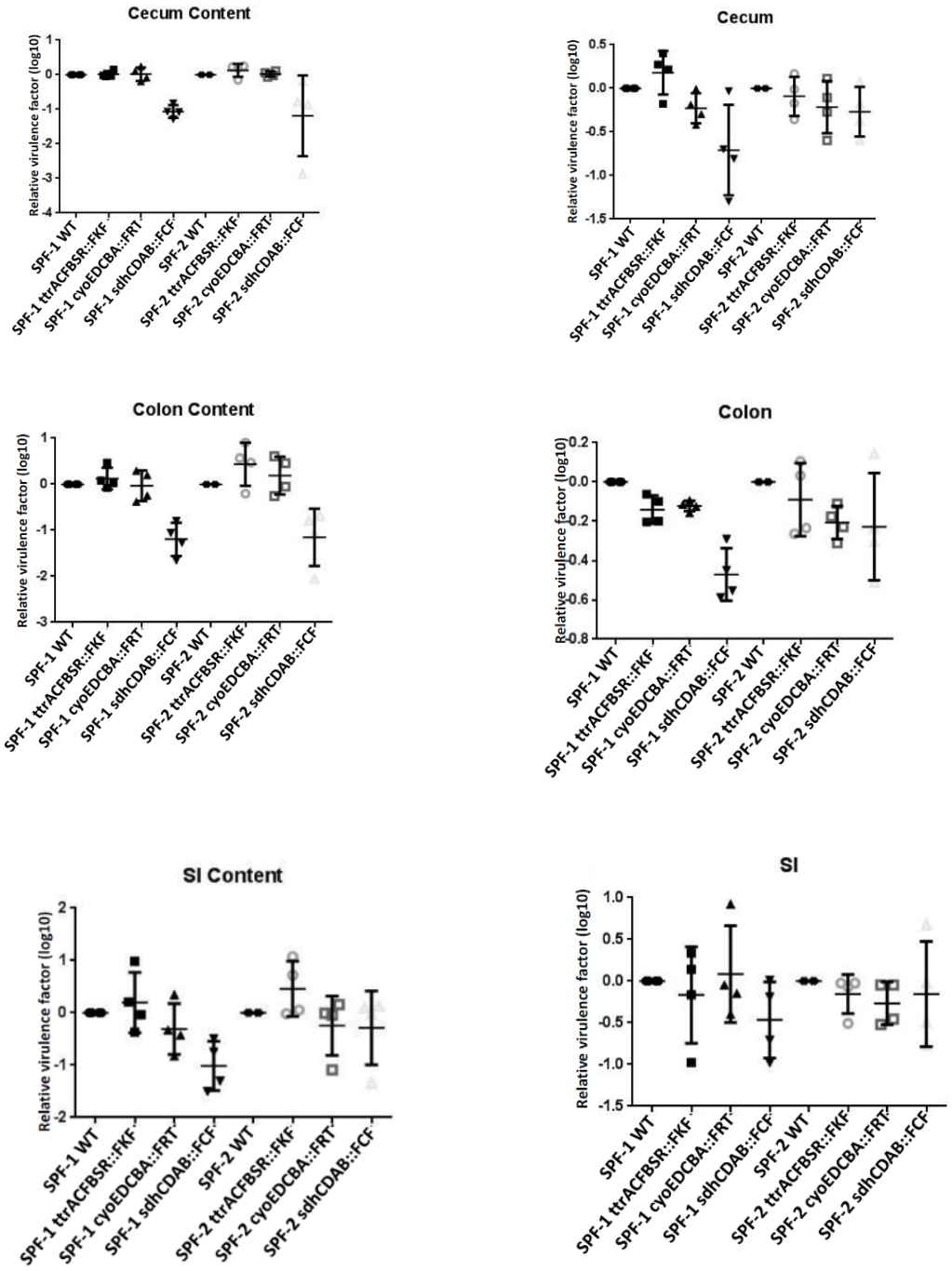
Figure 7-16 Differentially regulated nutrient uptake and utilization systems of S. Tm in

The picture depict the differential transcriptional regulatory networks employed by *Salmonella* under Germ-free (A) and colonized mouse models (B). All the genes mentioned above were observed to be differentially expressed or regulated, obtained from operon-based method by comparing Germ-free samples against those samples from colonized mouse models.

Germ-Free and colonized mouse models

From the non-coding RNA analysis, we also observed SgrS responsible for trafficking sugar-phosphates and maintaining the sugar-phosphate stress was highly expressed across all the *in vivo* conditions. This is depicted by **Figure 7-16**. Based on these observations, one could hypothesis that *S*. Tm is equipped with genetic elements that allow it to feed on multi-nutrient sources and possess regulatory elements controlling the sugar-phosphate stress, to produce energy *in vivo*. However, in the Germ-free condition, since there is no competition for nutrients, they feed on simple carbon source like citrate for energy production and cause infection.

7.5 Experimental validation



Relative count of viable *S*. Tm cells (ampicillin resistant) (CFU/g) from infected C57BL/6J mice from the small intestine, colon and cecum content and tissue (denoted by SPF-1 WT and SPF-2 WT) was compared with that of the count of *S*. Tm with ttr-operon mutant (denoted by ttrACFBSR), sdh-operon mutant (denoted by sdhCDAB), cyo-operon mutant (coyEDCBA) mutant created with antibiotic resistances for chloramphenicol (sdhCDAB), kanamycin (ttrACFBSR) or tetracyclin (cyoEDCBA). X-axis represent the *S*. Tm strains and Y-axis represent the logarithm difference between the CFUs found in wild type against those found in mutant strains.

Figure 7-17 Experimental validation

In order to evaluate whether the gene signatures observed in the sections above have any direct relationship with *S*. Tm infection, we picked 1) Nutrient uptake systems (like

the citrate and cellobiose uptake and utilization systems) and 2) Energy metabolism systems- cyo-operon involved in Complex I of oxidative phosphorylation, succinate dehydrogenase sdh-operon involved in oxidative phosphorylation and in TCA Cycle and tetrathionate uptake system ttr-operon that is an electron acceptor and involved in propanediol utilization system and created the *S*. Tm mutants of these operons. We observed only sdh-operon (encoding succinate dehydrogenase enzyme) mutant strains showed reasonable decrease in virulence in SPF-1 condition (in the cecum content and small intestine). This is depicted in **Figure 7-17** *Experimental validation*. This indicate that, a disturbance in complete utilization of TCA Cycle could affect *S*. Tm virulence in the SPF-1 condition.

7.6 *S*. Tm -Microbiota interactions

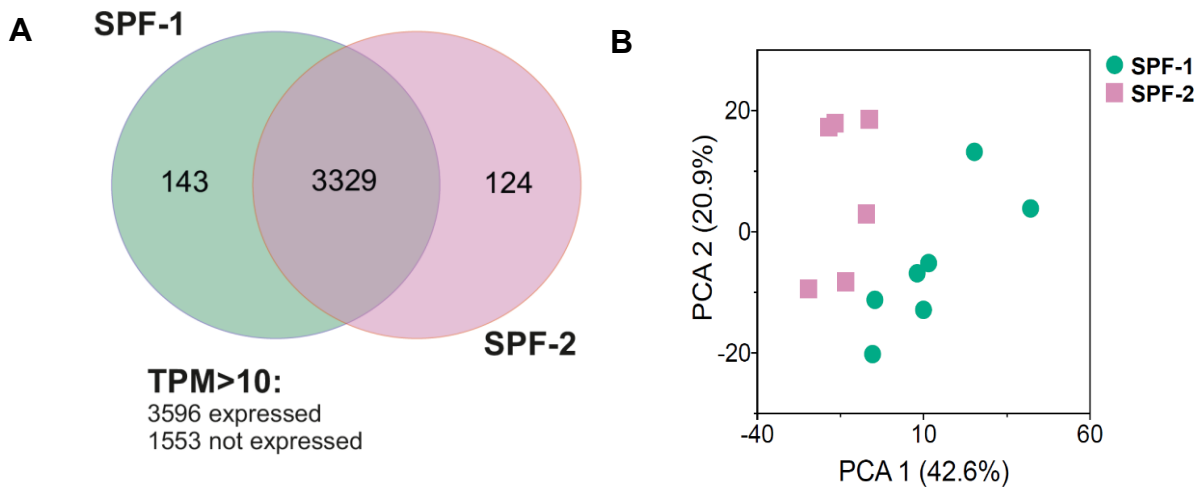


Figure 7-18 Comparative analysis of S. Tm expression in SPF-1 and SPF-2 conditions

A. Genes with an average TPM>10 were considered expressed for SPF-2 and SPF-1 conditions. The venn diagram represent the comparison of this list of genes expressed/not expressed for SPF-2 and SPF-1 conditions. The genes at the intersection represent those genes that were expressed in both SPF-2 and SPF-1 conditions. *B*. We obtained the ordination plot for *Salmonella* expression profile for all the samples from SPF-1 and SPF-2 conditions. The x-axis represent the Principal Component-1 that represent ~43% of the overall variance and Principal Component-2 that represent ~21% of the overall variance respectively.

In the previous section, we discussed in detail how *S*. Tm re-wired its transcriptional machinery in varied environments it encounter. In this section, we discuss on how microbiota influence the re-wiring of transcriptional machinery of *S*. Tm, with varied gut microbial compositions in the SPF-1 and SPF-2 conditions.

As mentioned in the previous sections, the SPF-1 gut microbiota environment is more susceptible to *S*. Tm infection in comparison to the SPF-2 gut microbiota environment (in cecal tissue). Hence, it is essential to identify the *S*. Tm transcriptional systems that were active under SPF-1 condition and its possible association with the gut microbiota. By doing so, we could identify the taxonomic groups and/or the functional groups associated with *S*. Tm virulence. Similarly, identifying the differentially active taxonomic groups and differentially expressed functionally groups could be helpful in identifying the microenvironments that aid in *S*. Tm infection.

In SPF-2 and SPF-1 *in vivo* conditions, *S*. Tm respond to a combinatorial effect of both the host gut environment and the gut-microbiota activity. In order to identify the potential *S*. Tm -microbiota interactions, we first performed an association study between the *S*. Tm expression profile and the metatranscriptome profile. The first section provides information on the differentially regulated *S*. Tm transcriptional machineries. The second section provides information on the relatively active taxonomic and functional units in the gut environment. Then, we identify the potential association between these signatures.

7.6.1 Comparing *S*. Tm expression profile in SPF-1 and SPF-2 conditions:

We performed the qualitative comparative analysis for *S*. Tm expression profile in SPF-2 and SPF-1 conditions. We considered genes with a median TPM value greater than 10 to be expressed in a particular condition (**Figure 7-18**-**A**). Based on this, we observed that around 3329 out of 5149 genomic features i.e. ~65 % expressed in both the SPF-2 and SPF-1 conditions. Only 143 genomic features (~2.4%) and 124 (~2.7%) of the overall genomic features were uniquely expressed in SPF-2 and SPF-1 conditions respectively.

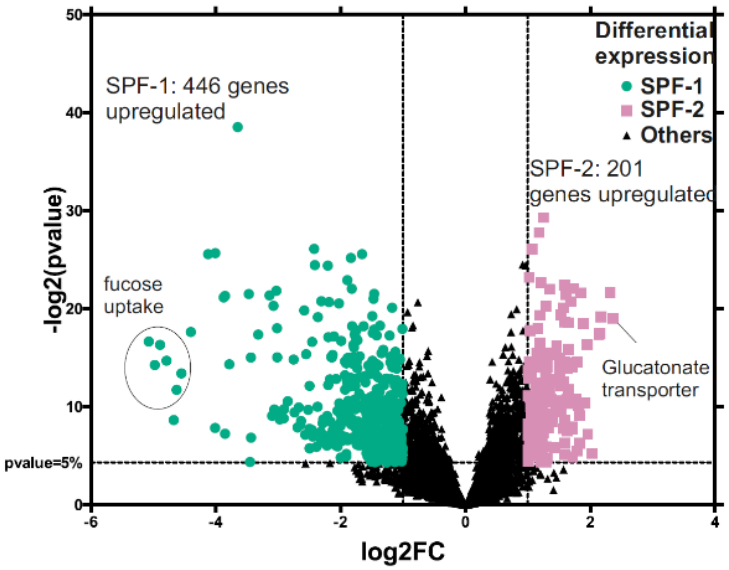


Figure 7-19 Volcano plot for *S*. Tm Differentially Expressed Genes between SPF-1 and SPF-2 conditions.

We obtained the *Salmonella* operons that were differentially expressed between the SPF-1 and SPF-2 conditions. Then, we picked the log2Fold Change and p-value obtained for those genes by DESeq2 and plotted the log2Fold Change in x-axis against the –log2 of the p-value in y-axis. The genes in red represent those genes highly expressed in SPF-2 condition and those in green represent those genes highly expressed in SPF-1 condition.

We observed 647 genes differentially expressed between the SPF-1 and SPF-2 conditions. Out of them, 446 genes were upregulated in SPF-1 condition and 201 genes were upregulated in SPF-2 condition. Amongst those that were upregulated in SPF-1 condition, many of them were involved in nucleotide, amino acid, coenzymes transport and utilization systems and energy production and conversion systems as shown in **Figure 7-19**.

Then, we looked for the differentially expressed TUs in this comparison. There were 253 unique and non-overlapping TUs (425 genes) that were upregulated in the SPF-1 condition against the SPF-2 condition and 126 unique and non-overlapping TUs (182 genes) that were upregulated in the SPF-2 condition against the SPF-1 condition. Here, we used the functional descriptions for the up and down-regulated genes from the supplementary data in (Ramachandran et al., 2012) to understand the difference in expression pattern between these two conditions.

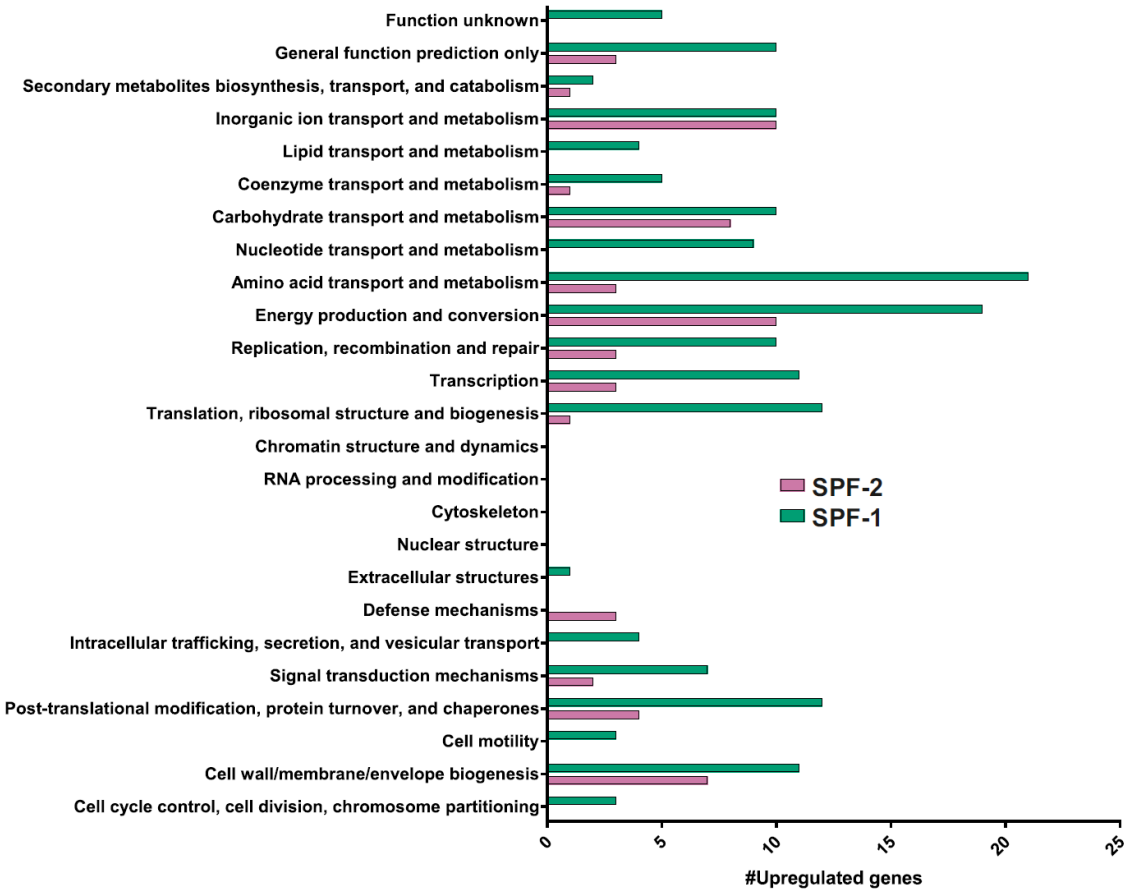


Figure 7-20 Functional groups (COG) upregulated in *S*. Tm in SPF-1 and SPF-2 condition

We obtained the genes that were differentially expressed in *S*. Tm expression profile from SPF-1 and SPF-2 conditions. We organized them into the COG functional category they belong. The histogram represents the number of upregulated genes for each functional category in both the SPF-1 and SPF-2 conditions.

7.6.1.1 *S*. Tm gene signatures in the SPF-1 condition:

Many transcriptional machineries associated with the carbon and nitrogen sources uptake/utilization systems and the virulence factors were upregulated under the SPF-1 condition, which explains the ability of *S*. Tm to out-compete other gut microbes to feed on simple nutrient sources and thereby the host's susceptibility to *S*. Tm infection.

1. Amongst the nutritional niches, the uptake/utilization systems associated with arabinose, cellobiose, fucose, rhamnose, sialic acid, glycerol-3-phosphate were upregulated in the SPF-1 condition. The propanediol uptake/utilization system and the corresponding master regulator *pocR* was also upregulated. Here, the cobalamine uptake and utilization-associated operons were also upregulated, which could indicate that the propanediol metabolism is happening in a cobalamine-dependent manner (Cheng and Bobik, 2010). The tetrathionate reductase enzyme that is associated with the growth advantage to overcome the competing microbiota in the mouse gut region (Hensel et al., 1999) was also

upregulated in the SPF-1 condition. In addition, a complete utilization of TCA cycle for energy production was also upregulated in SPF-1 condition. This is depicted in the **Figure 7-21**. There are few previous studies indicating the possible association between a complete utilization of TCA cycle and virulence in murine models (Tchawa Yimga et al., 2006, Bowden et al.,2010).

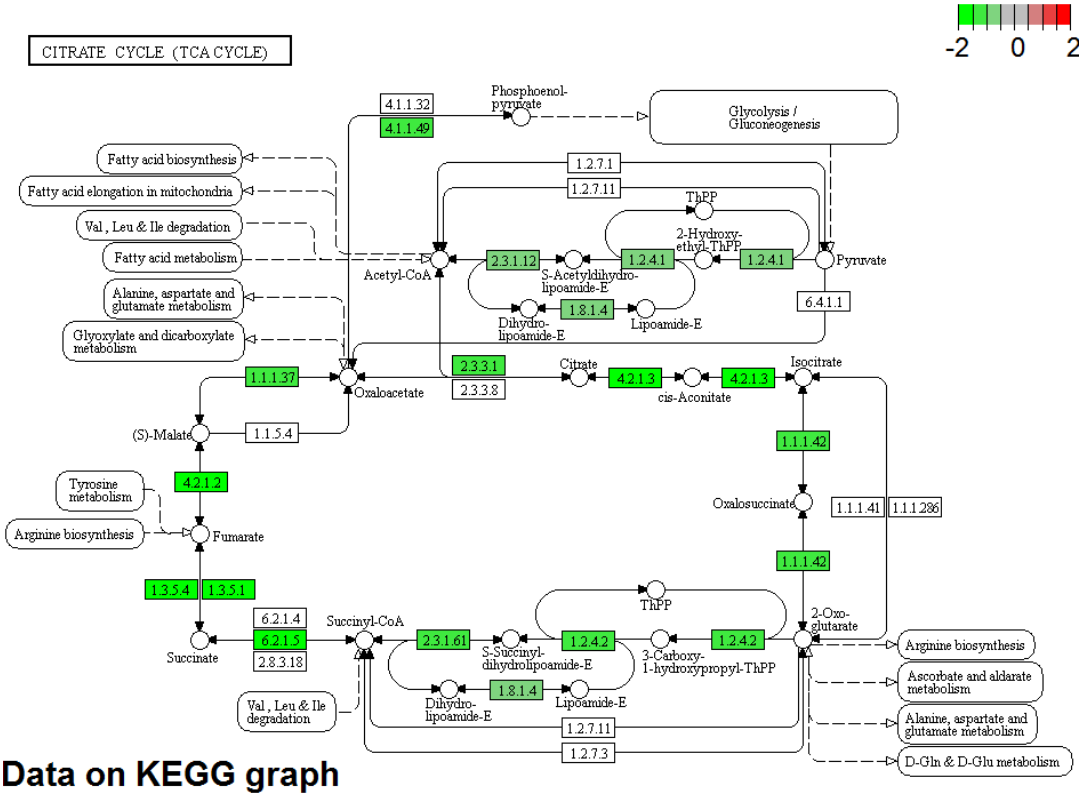


Figure 7-21 TCA cycle upregulation in *S*. Tm in SPF-1 condition

We mapped the logFold Change values of genes involved in TCA Cycle to the KEGG Pathway. The gene colored in green indicate they were highly expressed in the SPF-1 condition.

2. Electron transport chain: Complex II of the oxidative phosphorylation (succinate dehydrogenase) and cyo-operon which is known to be active under higher oxygen tension (Goldman et al., 1996) was upregulated in the SPF-1 condition. Also, phs-operon encoding oxidoreductase (conversion of thiosulfate to hydrogen sulfide) and ttr-operon (reduction of tetrathionate to thiosulfate) were upregulated in the SPF-1 condition; ttr-operon is known to provide competing

advantage over other microbiota in inflamed gut (Winter et al., 2010, Goldman et al., 1996).

3. Virulence factors *pagK* and *pagM* were also upregulated in *S*. Tm in the SPF-1 condition. The virulence plasmid pef operon responsible for cell adhesion was also upregulated in *S*. Tm.
4. Metal-ion transport: Iron storage proteins like bacterioferritin and ferritin-like protein were upregulated in the SPF-1 condition. The cob-operon and cbi-operon that were associated with cobalamine biosynthesis during anaerobic conditions and *S*. Tm infection were upregulated in the SPF-1 condition in *S*. Tm (Vaz et al., 2011).

In addition, genes that were associated with stress responses, polyamine metabolism, anaerobic metabolism and central intermediary metabolism were upregulated. Thus, one could infer that, under SPF-1 condition, *S*. Tm prefer host-derived carbon sources, utilize TCA cycle for energy production and activate virulence factors.

7.6.1.2 *S*. Tm gene signatures in the SPF-2 condition:

Under SPF-2 gut microbiota condition, the host gut environment (cecal tissue) is less susceptible to *S*. Tm infection. When compared to the SPF-1 condition, *S*. Tm has far less number of genes involved in central intermediary metabolism, stress responses and polyamine metabolism upregulated in the SPF-2 condition.

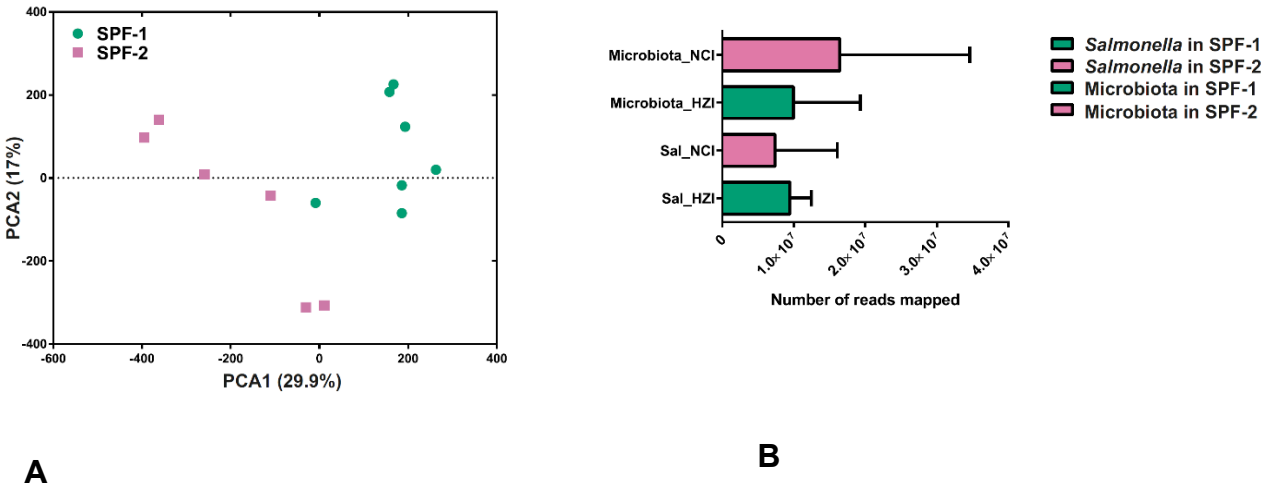
1. Operons encoding 30S, 50S ribosomal binding proteins that were involved in translation or protein modification was upregulated in *S*. Tm in the SPF-2 condition.
2. Electron transport chain: Puridine and pyrimidine metabolism associated transcriptional units were upregulated in *S*. Tm in the SPF-2 condition. The Rnf electron transport complex which is known to obtain electrons from reduced ferrodoxin and utilize it in the electron transport system (Schlegel et al., 2012) and possibly couple them to the pyrimidine nucleotide pool (Biegel et al., 2011) was also upregulated in *S*. Tm in the SPF-2 condition.
3. Nutritional niches: *S*. Tm upregulated genes to uptake and utilize glucose derivatives (gluconate, galactose) and C4-dicarboxylates in the SPF-2 condition.
4. Metal-ion transport: The cor-Operon responsible for magnesium/nickel transport was highly expressed in the SPF-2 condition.

5. The operon that encodes O-antigen capsules (Group IV polysaccharide capsules) which is associated with the host immune system evasion (Marshall and Gunn, 2015), was upregulated in the SPF-2 condition.

The upregulation of this compendium of transcriptional machineries indicate that under SPF-2 condition, *S*. Tm intend to survive by feeding on multiple nutrient sources and downregulating the TCA cycle and the virulence genes. The difference between the two murine models could be attributed to their difference in gut microbial composition and this difference in microbiota composition created different microenvironments that *S*. Tm encounter, thus playing a strong role in modulating the *S*. Tm infection.

7.6.2 Microbial signatures

The previous section discussed the different functional systems employed by *S*. Tm to encounter the microenvironments created by the different gut microbial communities in the SPF-1 and SPF-2 condition. In this section, we identify the potential microenvironments encountered by the *S*. Tm in the SPF-1 and SPF-2 condition. We infer them, by identifying the active members of the community and the corresponding functional systems active from the metatranscriptome data.



A **B**

Figure 7-22 SPF-1 and SPF-2 metatranscriptome read map statistics

A. The figure represents the ordination plot obtained for the expression profile of the SPF-1 and SPF-2 gut microbiota. The x-axis represent Principal Component-1 and y-axis represent Principal Component-2 constituting ~30% and 17% of the overall variance respectively. B. The figure represents the barplot for the number of reads mapped to Salmonella and the gut catalogue reference in SPF-1 and SPF-2 conditions with errorbars representing one std.deviation.

In order to characterize the composition and activity for the microbial communities competing with *S*. Tm during infection of the SPF-1 and SPF-2 mice, we compared the metatranscriptome during infection. Meta-transcriptome profile in the SPF-1 and SPF-2 condition were mapped against a mouse microbiota gene catalog containing functional and taxonomic annotations developed by Lesker et al. The brief outline of the workflow employed to obtain the metatranscriptome profile is given in **Figure 7-23**. Mapping statistics for each samples are depicted in **Figure 7-22-B**.

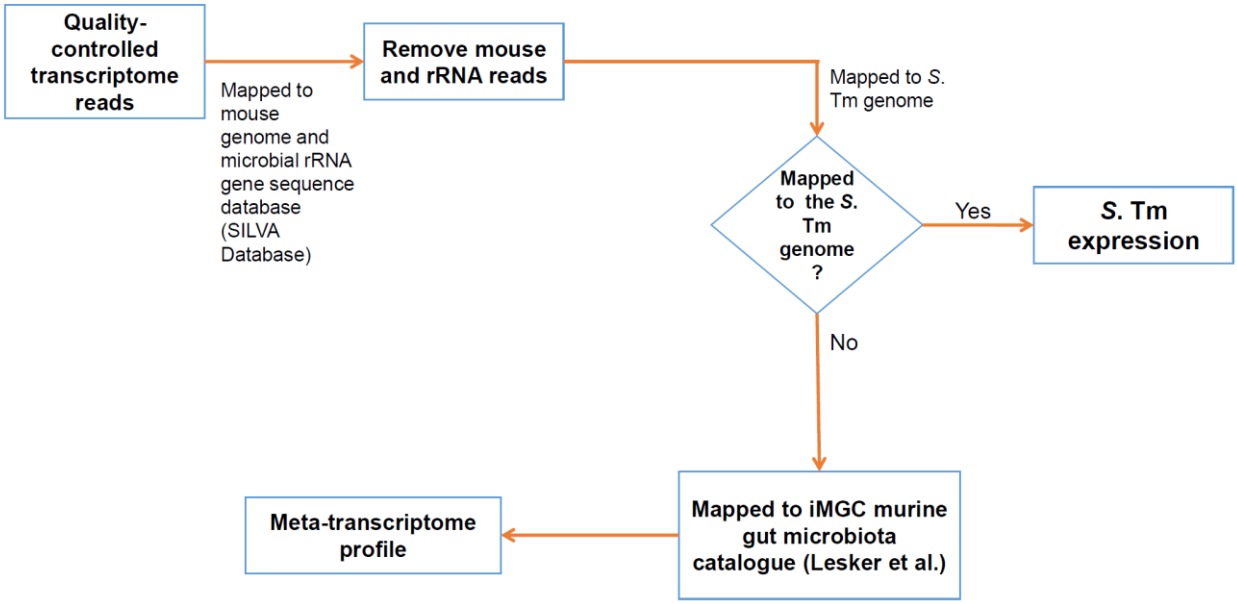


Figure 7-23 Brief outline of the workflow employed to obtain the metatranscriptome profile

The initial mapping statistics of the meta-transcriptome reads showed that both the SPF-1 and SPF-2 samples had, on average around 11 and 15 million reads respectively whereas there were around 8 million and 7 million reads mapped to the *S*. Tm under the SPF-1 and SPF-2 conditions respectively, indicating the samples has sufficient coverage to carry out the metatranscriptome study. The ordination plot of the meta-transcriptome data (**Figure 7-22-A**) give us an indication that the SPF-1 gut environment had distinct active members and functional units when compared to that of the SPF-2 samples.

7.6.2.1 Active members of the community:

The active members and / or the active functional components of the gut microbiota individual *in vivo* models were distinct for each condition. To investigate this, we identified the active members of the SPF-1 and SPF-2 gut microbiota community. For this, we mapped the reads that do not belong to either mouse or *S*. Tm genome to a mouse gut microbiota catalogue to identify the active members of the microbiota community in the colonized mouse models SPF-1 and SPF-2 conditions. After mapping and quantifying the number of reads mapping to a specific taxonomic bin, we normalized it to the scale of transcripts per millions. We considered those taxonomic units with the mapped-reads percentage of 0.5% (at least 5000 TPM) to be active. From these profiles, we were interested in identifying a) The active members of the gut microbial communities SPF-1 and SPF-2 b) the activity profile of the gut microbial community at different taxonomic level. We observed there were 13 active members in the SPF-1 condition and 18 active members in the SPF-2 condition. The active members of the SPF-1 and SPF-2 gut microbial communities constitute ~62% and ~52% of the overall metratranscriptome profiles respectively. To observe the activity profile at different taxonomic levels, we organized the taxonomic units to the taxonomic levels (from phylum to species) they belong. This is depicted below in **Figure 7-24** and discuss in detail in this section.

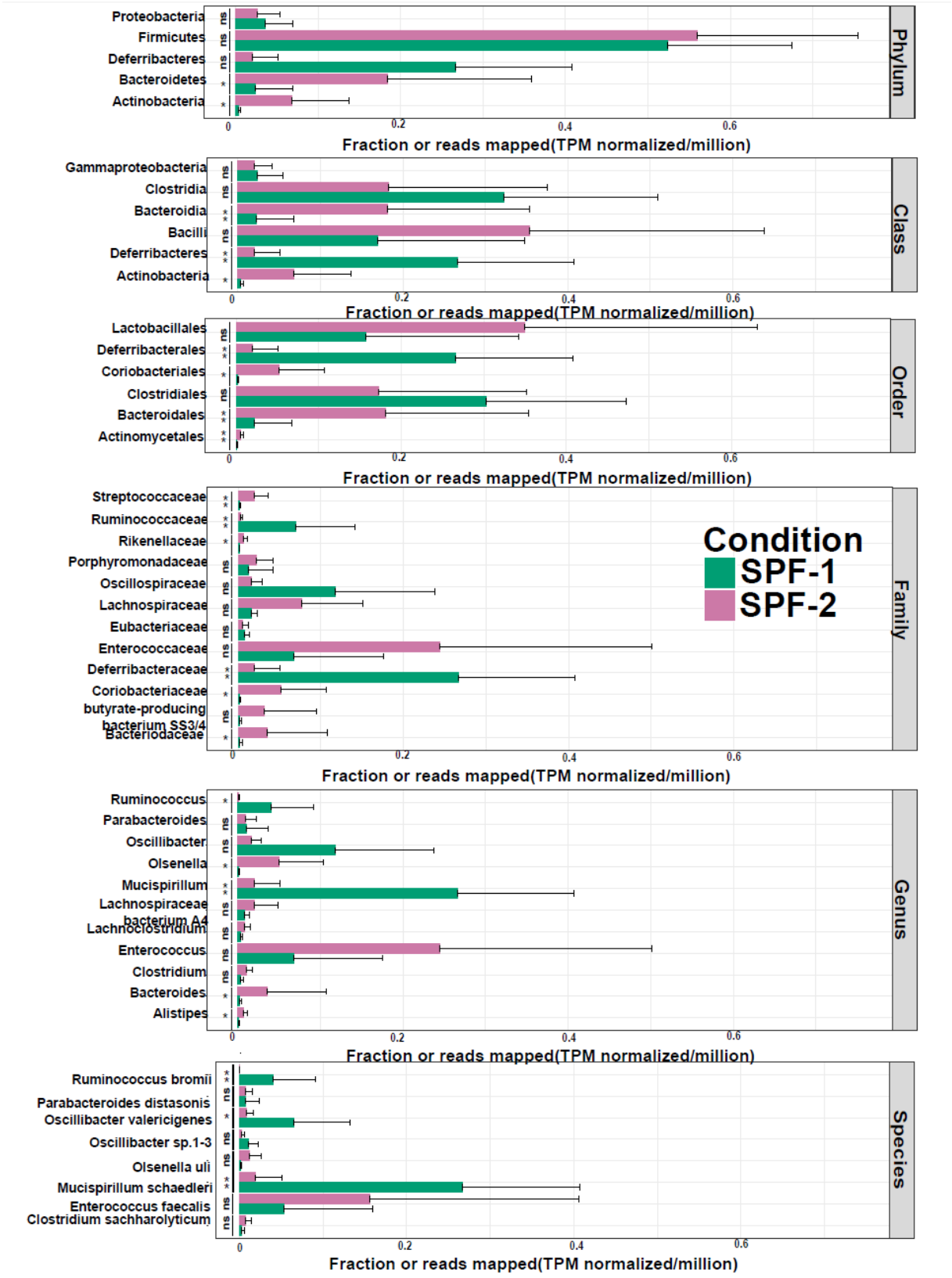


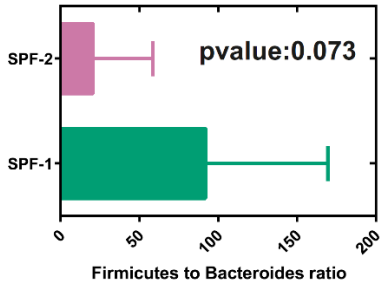
Figure 7-24 Activity profile of SPF-1 and SPF-2 gut microbial community at various

The barplot represents the fraction of reads (TPM normalized) mapped to different taxonomic groups. The x-axis denote the fraction of reads and y-axis denote the taxonomic groups. They were segregated into the taxonomic level they belong. The red bars and green bars represent the fraction of reads mapped to a taxonomic group in SPF-2 and SPF-1 gut microbial communities respectively. The difference in activity of each taxonomic group between SPF-1 and SPF-2 gut microbial communities is indicated by the asterisks provided near the bars (ns: not significant, *-pval<5%), **-pval<1%, ***-pval<0.001, ****-pval<0.0001

taxonomic levels

7.6.2.1.1 Phylum:

In SPF-2 gut microbial community, microbes of phyla *Bacteroidetes* and *Actinobacteria* were active when compared to the SPF-1 gut microbial community and those of phyla *Deferribacteria* were active in the SPF-1 gut microbial community in comparison to the SPF-2 gut microbial community. *Firmicutes* were highly active in both the SPF-1 and SPF-2 gut microbial communities. When we compared the *Firmicutes* to *Bacteroidetes* Ratio, it was very high in the SPF-1 gut microbial community (mean=~90) when compared to SPF-2 gut microbial community (mean=~20) as shown below. It has been previously reported an association between increased *Firmicutes* to *Bacteroides* ratio and susceptibility to obesity (Ley et al., 2005, Ley et al., 2006, Koliada et al., 2017) and other diseases. The reduction in *S*. Tm infection (in cecal tissue) for the SPF-2 gut microbial community in comparison to the SPF-1 gut microbial community could be associated with a reduction in the *Firmicutes* to *Bacteroides* ratio.



We obtained the ratio between the number of reads mapped (TPM normalized) to Phyla *Firmicutes* and Phyla *Bacteroides* and the bars represent the mean values with errorbars for one standard deviation. A Wilcoxon rank-sum test between the ratios obtained for SPF-1 and SPF-2 gut microbial communities was around 8%.

Figure 7-25 *Firmicutes* to *Bacteroides* ratio for the SPF-1 and SPF-2 gut microbial communities

7.6.2.1.2 Order:

The order-level taxonomic assignments reveal that *Clostridiales* and *Lactobacillales*, the common inhabitants of the gut microbiota ecosystem were active in both the SPF-1 and SPF-2 gut microbial communities.

7.6.2.1.3 Family:

Around 55% % of the SPF-2 and SPF-1 meta-transcriptome reads mapped to the gut microbiota catalogue belonged to family taxonomic units. *Deferribactericae* was relatively active in the SPF-1 gut microbial community and *Enterococcaceae* was relatively active in the SPF-2 gut microbial community.

It is interesting to note that in SPF-1 condition that is more susceptible to *S*. Tm infection in cecal tissue, *Deferribacteria* was comparatively very active when compared to that in the SPF-2 condition. *Deferribacteria* is earlier reported to be highly abundant

in *Enterobacteriaceae C. rodentium* infection (Hoffmann et al., 2009). Similarly, Enterococcus bacteria like *Enterococcus faecelis* was active in both SPF-2 and SPF-1 condition.

7.6.2.1.4 Genus:

Around 50% of the SPF-1 and SPF-2 meta-transcriptome reads mapped to the gut microbiota catalogue belonged to genus taxonomic units respectively. Species belonging to *Oscillibacter* were active in the SPF-1 gut microbial community when compared to the SPF-2 gut microbial community. These microbes, along with *Ruminococcus* were observed to be highly active in individuals fed with Resistant Starch (RS) diet (Clarke et al., 2012). This starch is hard to digest and absorb by the host digestive system since it is dominated by straight-chain polysaccharides. When broken down, they yield short-chain fatty acids.

7.6.2.1.5 Species:

Around 40% of the SPF-2 and SPF-1 meta-transcriptome reads mapped to the gut microbiota catalogue belonged to species. *Mucispirillum schaedleri* and *Ruminococcus bromii* were the top active members of the SPF-1 gut microbial community. *Mucispirillum schaedleri*, a known pathobiont commonly observed in mouse gut and occasionally in human gut mucus region, is studied to possess genes to utilize, not degrade host-glycans (Loy et al., 2017). *Ruminococcus bromii* is studied to degrade resistant starch sources present in human colon (Ze et al., 2012). *Enterococcus faecelis* and *Olsenella uli* (lactic-acid related bacteria) were the top active members of the SPF-2 gut microbial community.

We observed there were 13 active members in SPF-1 and 18 active members in SPF-2 condition. This is represented in **Figure 7-26** for SPF-1 gut microbiota and **Figure 7-27** for SPF-2 gut microbiota.

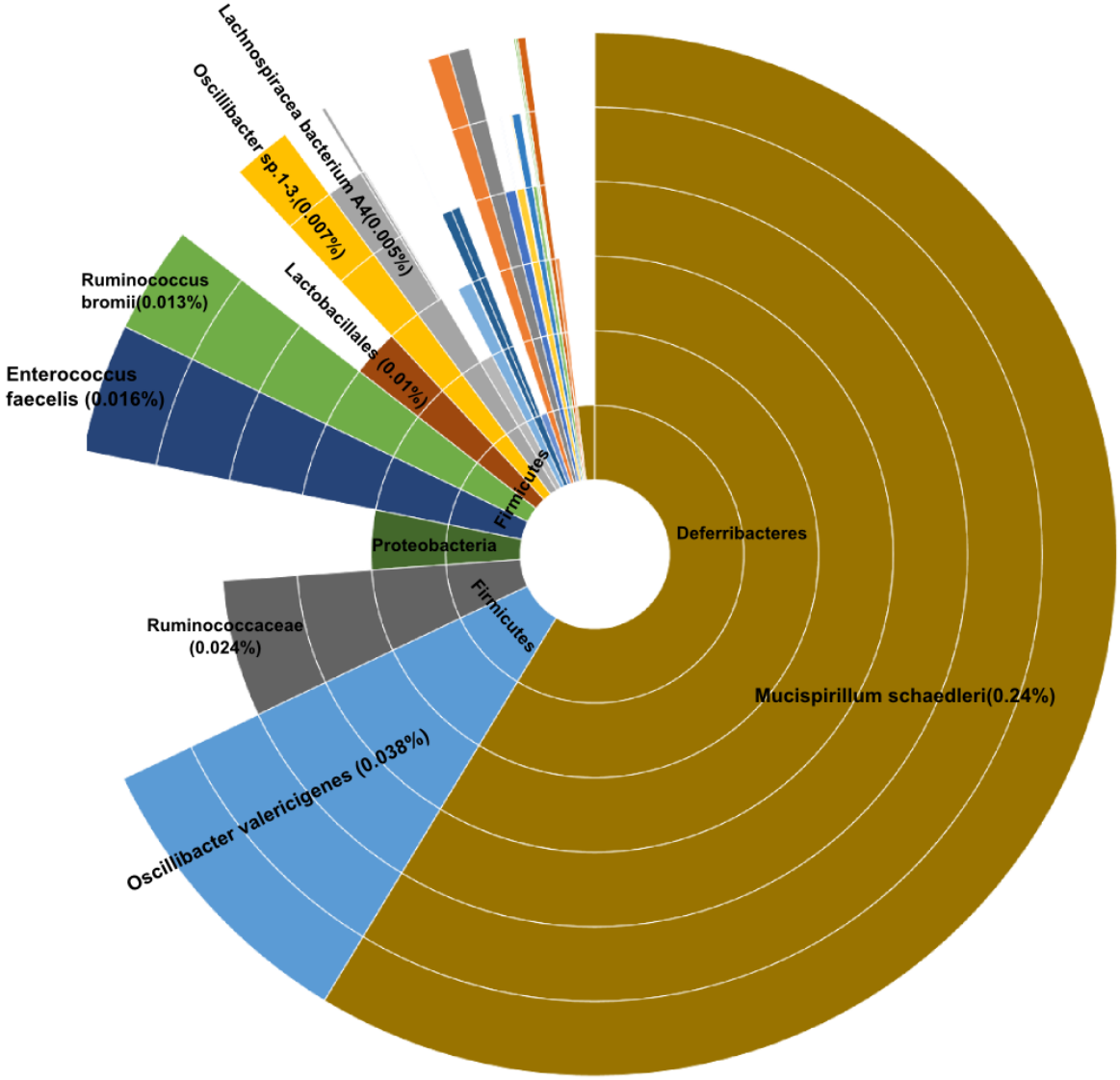


Figure 7-26 Sunburst plot representing the taxonomic diversity of the active members of the SPF-1 community.

The different levels of the concentric circles of the sunburst diagram represent the different levels of taxonomic assignment to the gut microbial community and the area represent their relative activity. In addition, these sunburst diagrams indicate the efficiency of the mouse gut microbiome catalogue in providing species-level taxonomic assignments for the active members of the community.

The sunburst diagram of the relative activity of the members of the SPF-1 and SPF-2 community reveal that they encompass distinct active members. In SPF-1 condition, in spite of their high-complexity, few species were highly active belonging to *Deferribacteria* phyla namely *Mucispirillum schaedleri* and *Firmicutes* phyla namely *Enterococcus faecelis*, *Ruminococcus bromii* and *Oscillibacter valericigenes*.

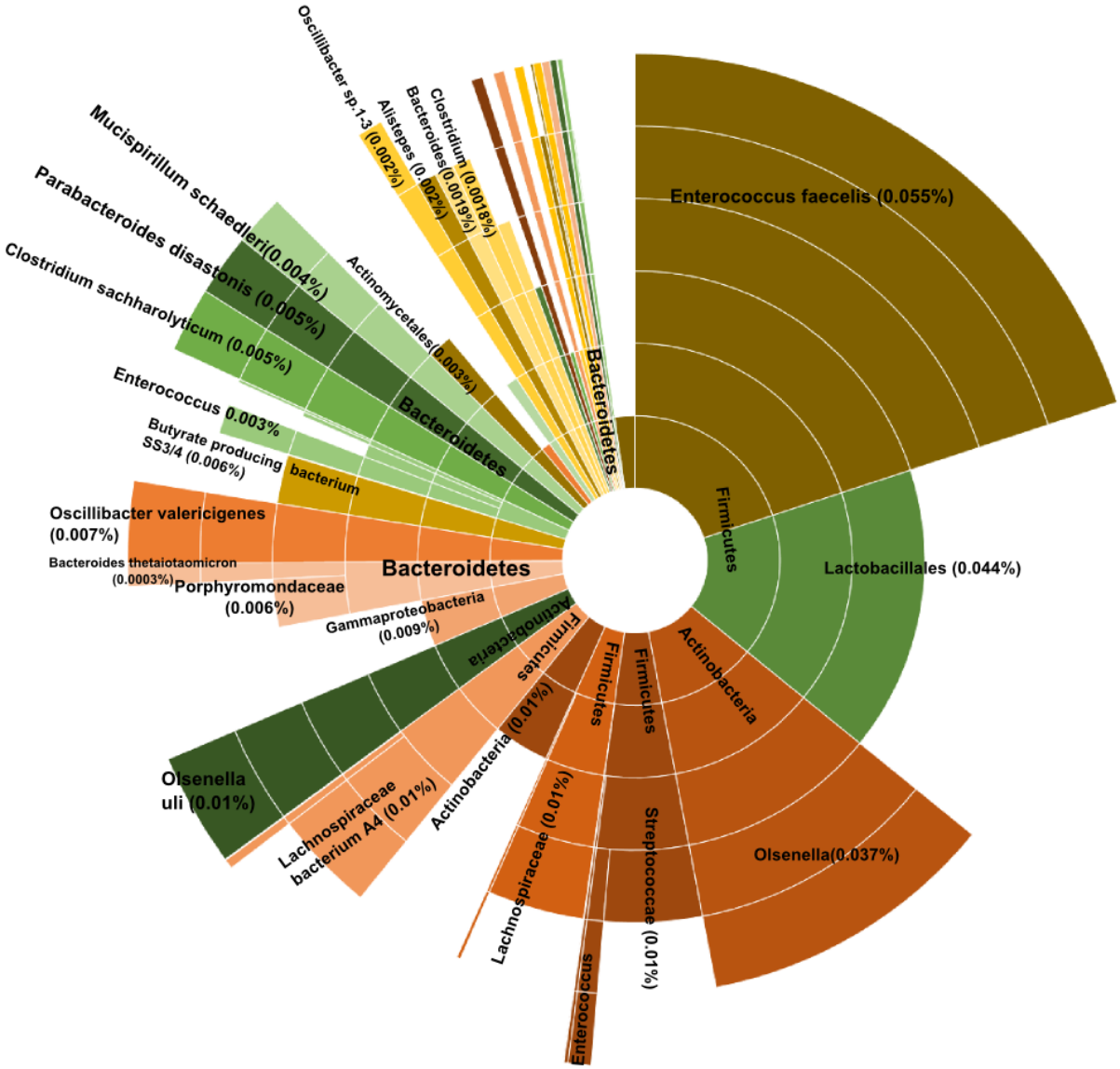


Figure 7-27 Sunburst plot representing the taxonomic diversity of the active members of the SPF-2 community.

In SPF-2 condition, few species belonging to phyla *Firmicutes* namely *Enterococcus faecelis* and *Lachnospiracea bacterium A4* and phyla *Actinobacteria* namely *Olsenella uli* were highly active. *Deferribacteria* is earlier reported to be highly abundant in *Enterobacteriaceae C.* rodentium infection (Hoffmann et al., 2009). Similarly, *Enterococcus* bacteria like *Enterococcus faecelis* reported earlier to provide protection against *S.* Tm infection (Rangan et al., 2016). Even though these microbes were active in both the gut microbial environments, they were unable to provide a complete protection against *S.* Tm in both the SPF-1 and SPF-2 conditions.

7.6.2.2 Differentially active taxonomic groups:

Similar to identifying the active members of the gut microbial microbiota, we obtained those microbes that were differentially active. For this, we calculated the sum of the expression values of all the genes belonging to each taxonomic group and observed the differentially expressed taxonomic groups in the SPF-1 and SPF-2 gut microbial communities. We plotted those taxonomic groups that obtained a log2FoldChange greater than ±3 and adjusted p-value less than 5e-02. As shown in **Figure 7-27**, there were 31 highly differentially active taxonomic units. Amongst them, there were around 17 taxonomic units comparatively active in the SPF-2 gut microbial community. The relatively active microbiota community of the SPF-2 gut microbial community was more diverse when compared to that of the SPF-1 gut microbial community. It is interesting to note that in SPF-1 condition (which is highly susceptible to *S*. Tm infection), *Bacteroides* were

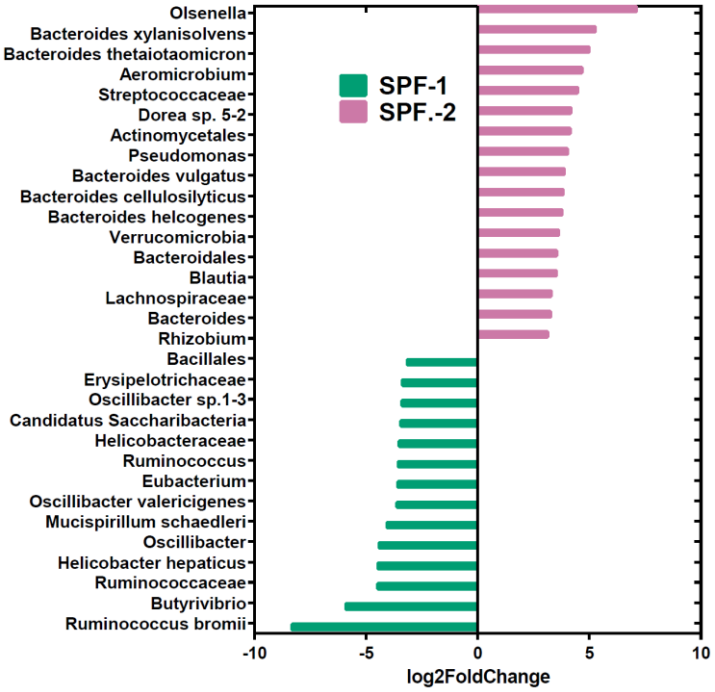


Figure 7-28 Differentially active members of the SPF-1 and SPF-2 gut microbial communities

We obtained the differential activity level of all the taxonomic groups using the number of reads mapped to them, via DESeq2. The x-axis denote the log2Fold Change of the taxonomic groups. The green bar indicate the corresponding taxonomic groups that were highly active in the SPF-1 gut microbial community and the red bar indicate the corresponding taxonomic groups that were highly active in the SPF-2 gut microbial community.

expressed lesser than that of the SPF-2 condition. There were 14 differentially active taxonomic units in the SPF-1 gut microbial community. *Proteobacteria* and *Deferribacteres* groups of bacteria represent the dominant members of the differentially active taxonomic units in the SPF-1 gut microbial community. Many of the

differentially active taxonomic groups in SPF-1 condition belonged to *Proteobacteria* and *Firmicutes* group while *Bacteriodetes* in SPF-2 condition. To be specific, few of the taxonomic groups that were studied to enhance the host defense against *S*. Tm infection- *Enterococcus faecium* (Rangan et al., 2016) is differentially active in the SPF-2 gut microbial community. This is shown in **Figure 7-27**.

7.6.3 Functional signatures

In the previous section, we identified the dominant members of the SPF-1 and SPF-2 gut microbial communities, who were potentially responsible for creating microenvironments that could be associated with *S*. Tm infection. In this section, we identify and study the functional signatures that could be associated with the microenvironments encountered by *S*. Tm. In order to do this, firstly we identified the potentially active functional components by two ways: one by comparing the number of genes that were differentially expressed in COG functional groups between SPF-2 and SPF-1 conditions. Next, we organized the differentially expressed genes into the KO groups they belong and determined the over-represented KEGG Modules and KEGG pathways.

7.6.3.1 Active COG functional groups in microbiota community:

We obtained the genes with a fold change greater than 2 and adjusted p-value≤ 1% and grouped them into the COG functional/sub-functional groups they belong. The bar plot 3.6M depict this comparison. Amongst them, SPF-1 microbiota possessed more number of genes upregulated in all the COG functional categories when compared to the SPF-2 microbiota. However, the functional category cell motility had distinct set of genes highly expressed in both the SPF-1 and SPF-2 gut microbial communities (**Figure 7-28**). Many genes that were differentially expressed do not possess a COG annotation, hence this do not provide the complete picture. To have a better perspective of the functional signatures, we performed a similar analysis based on KEGG annotations and this is discussed in the next section.

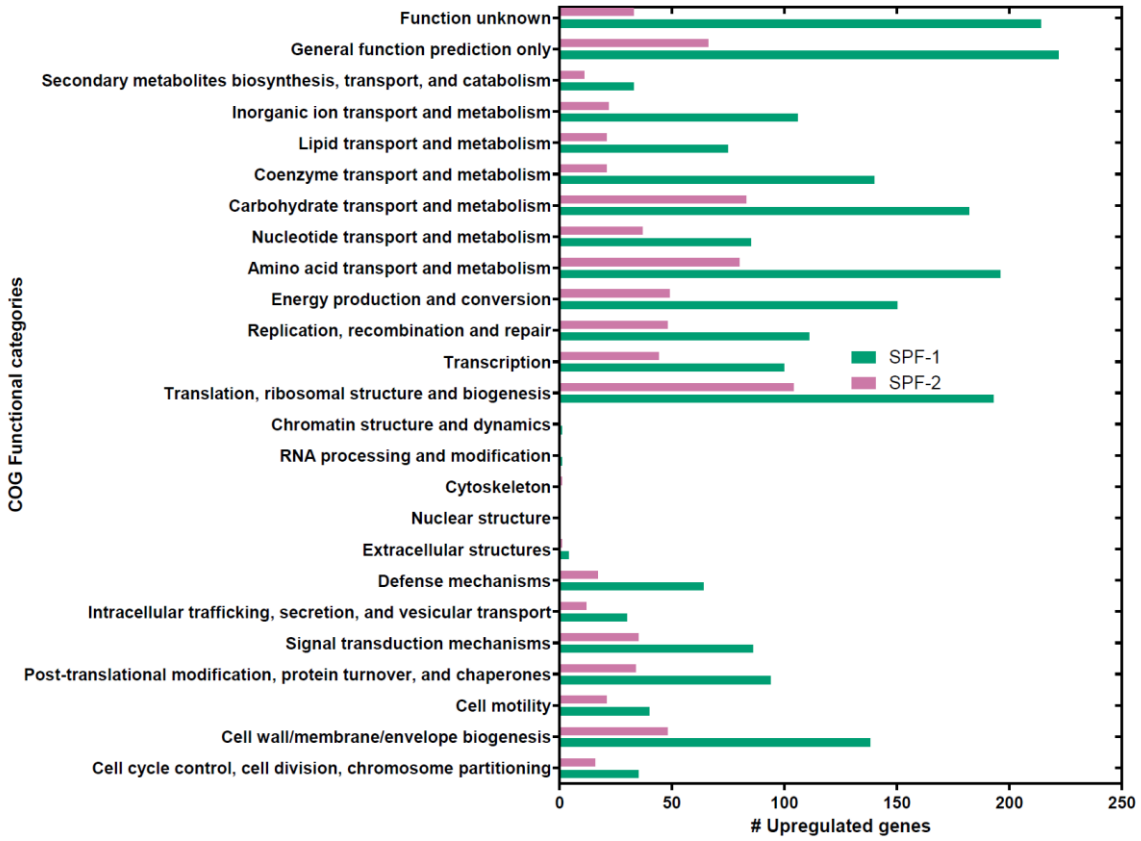


Figure 7-29 Functional groups (COG) upregulated in SPF-1 and SPF-2 gut microbial

We obtained the genes that were differentially expressed in SPF-1 and SPF-2 microbiota comparison. They were organized into the COG functional categories they belong. The histogram represents the number of upregulated genes in COG functional categories. The green bars and red bars indicate the number of genes upregulated in SPF-1 and SPF-2 condition belonging to a COG functional category

communities

7.6.3.2 Active KO functional modules and pathways in microbiota community:

In order to identify the active functional modules and pathways identified in the meta-transcriptome profile of the colonized mouse models, we removed the reads that belonged to mouse or *Salmonella* and mapped against the gut microbiota catalogue with KEGG Orthology annotation. We picked the KEGG orthologies with a fold change greater than or equal to 2. We identified the KEGG modules and pathways that were over-represented (Over-representation analysis with p-value <=5% for modules and 0.001% for pathways with atleast 40% of genes differentially expressed) in these selected KEGG Orthologies. Thus, we obtained 16 KEGG Pathways and 42 KEGG Modules upregulated in the SPF-1 condition. Correspondingly, there were 36 KEGG Pathways and 28 KEGG Modules upregulated in SPF-2 (**Figure 7-30**, **Figure 7-31**). Amongst them, we identified interesting pathways and modules that could be

associated with *S*. Tm transcriptional machinery in the respective gut microbiota conditions.

- **Bacterial Secretion System:** This KEGG pathway was upregulated in the SPF-1 microbiota. In this pathway, module involved in Type-II Secretion System was upregulated in the SPF-1 gut microbiota environment. This system translocates enzymes like proteases and chitinases that could aid *S*. Tm in feeding on cellobiose, sialic acid (Sandkvist, 2001), (Cianciotto and White, 2017). Correspondingly, cellobiose and sialic acid uptake systems were upregulated in *S*. Tm in SPF-1 condition.
- **Flagellar assembly and bacterial chemotaxis:** Out of 28 genes involved in flagellar assembly, 22 genes were highly expressed in SPF-1 condition. This include genes encoding for rotation of flagellar motors- *motA* and *motB*.
- **TCA Cycle:** The genes involved in citrate cycle were upregulated in both *S*. Tm and the corresponding gut-microbiota in the SPF-1 condition. This could indicate a competition for nutrient sources. In spite of this, the gut microbiota environment in SPF-1 condition provides favorable environment for *S*. Tm to feed on host-associated, microbiota-degraded carbon sources like cellobiose and sialic acid (as discussed before).
- **Oxidative phosphorylation**: The genes constituting F-type ATPase, the complex V of the oxidative phosphorylation were highly expressed in SPF-1 condition.
- **Cobalamine biosynthesis**: We observed genes responsible for cobalamine biosynthesis were upregulated in SPF-1 condition. Correspondingly, the cobalamine uptake and utilization system in *S*. Tm were upregulated in SPF-1 condition when compared to the SPF-2 condition. There have been previous studies observing *S*. Tm genus is avirulent in cob operon mutant strains (de Paiva et al., 2009). This indicate that *S*. Tm pathogenicity in SPF-1 condition might be associated with cobalamin biosynthesis by the gut microbiota and the corresponding cobalamine uptake and utilization by *Salmonella*. In addition, the modules responsible for metal-ions transport systems like manganese, zinc, nickel were upregulated in the SPF-1 condition. These evidences suggest that the dysbiotic SPF-1 community was able to provide the required metabolic niches for *S*. Tm to survive and cause virulence, in comparison to the SPF-2 gut microbial community.

- **Pentose Phosphate Pathway:** In SPF-2 condition, the Pentose Phosphate Pathway (20 out of 24 KOs differentially expressed) and glycolysis (25 out of 43 KOs differentially expressed), responsible for energy production and nucleotide synthesis from glucose was upregulated.
- In SPF-2 condition, the nicotinate and nicotinamide metabolism (12 out of 19 KOs differentially expressed) were highly expressed in comparison to the SPF-1 condition. They include 3 modules involved in aspartate to NAD conversion.
- **Oxidative Phosphorylation:** In SPF-2 condition, the V-type ATPase, which is known to hydrolyze ATP for proton transport across the cell membranes, was upregulated.

This *S*. Tm -microbiota association study give us an indication that the choice of nutrient for *S*. Tm in SPF-1 and SPF-2 conditions could be attributed to varied gut microbial compositions. For instance, in the SPF-1 gut microbial community, the gut microbiota synthesize cobalamine (cofactor for host-derived propanediol utilization) and *S*. Tm utilize them for its adaptation and virulence. Whereas in the SPF-2 gut microbial community, the association study cannot identify many possible interactions between *S*. Tm and the gut microbiota because many *S*. Tm genes upregulated in the SPF-2 condition were not annotated (genes with unknown functions). To identify more potential *S*. Tm -microbiota functional interactions, we introduce a novel approach that compares the expression profile of both *S*. Tm and microbiota from same gut microbiota.

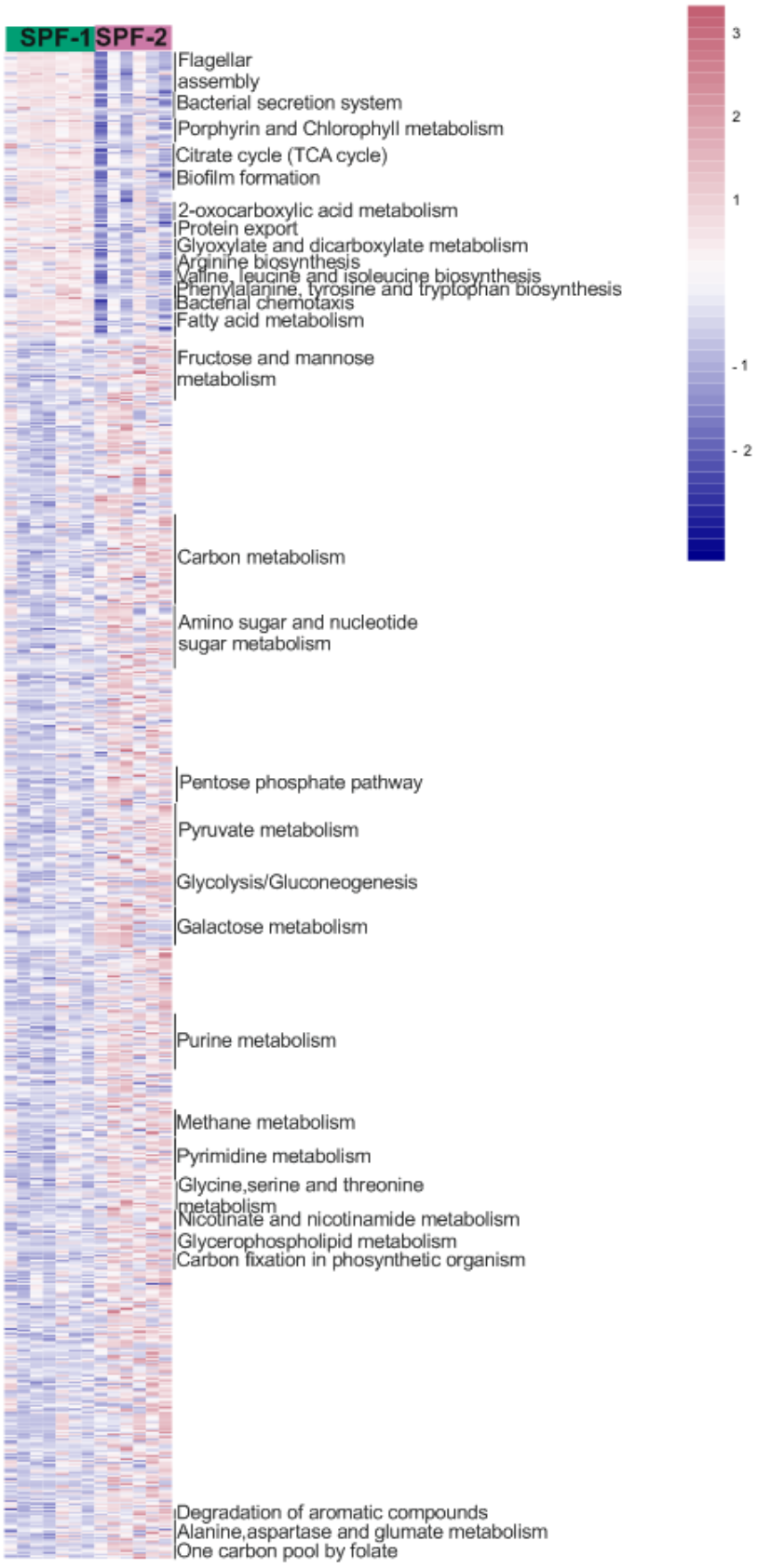
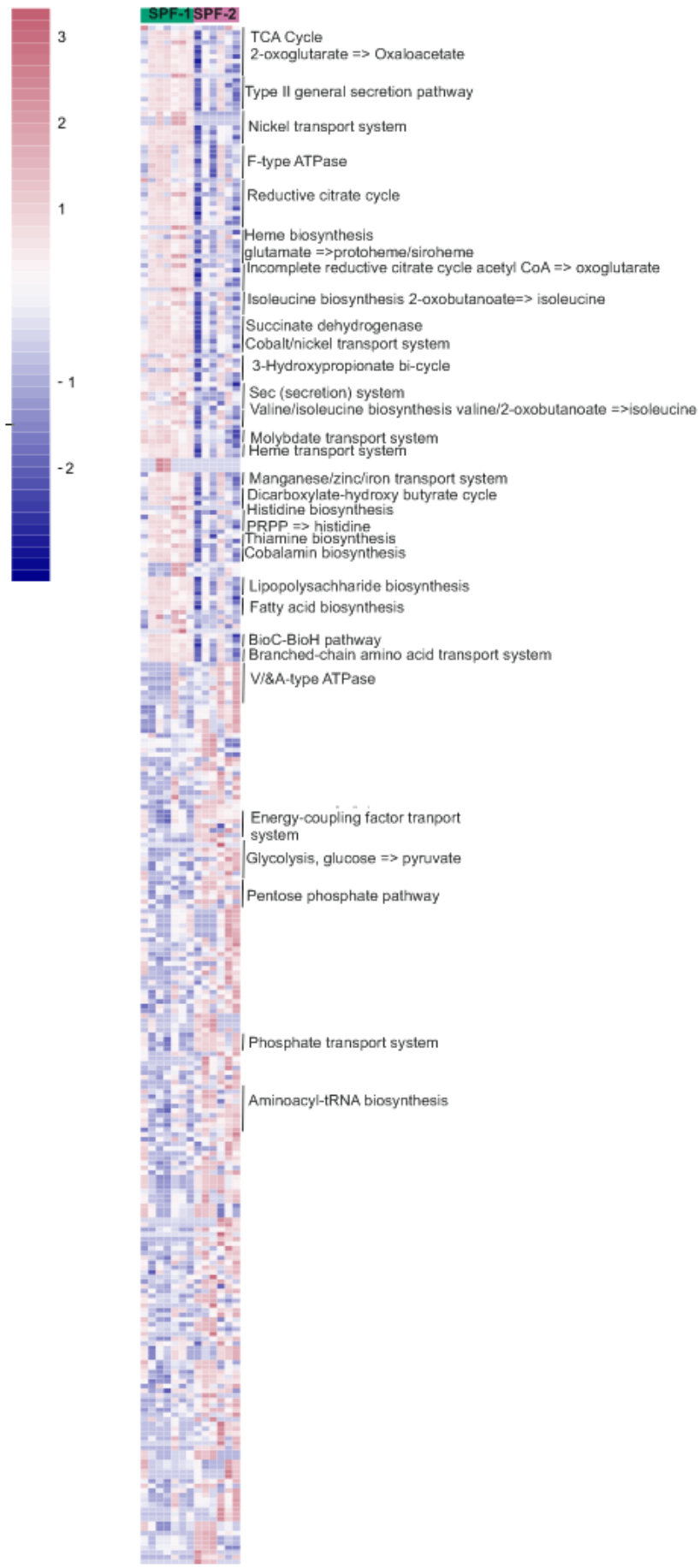


Figure 7-30 Functional groups (KEGG Pathways) differentially expressed in SPF-1 and SPF-2 gut microbial communities

We obtained those KEGG Orthologies that were differentially expressed in the SPF-1 and SPF-2 environment, organized them into the KEGG Pathways they belong. We enriched these KEGG Modules and Pathways by over-representation analysis. The heatmap represents these KEGG Pathways.



We obtained those KEGG Orthologies that were differentiallly expressed in the SPF-1 and SPF-2 environment, organized them into the KEGG Modules they belong. We enriched these KEGG Modules by over-representation analysis. The heatmap represents these KEGG Modules.

Figure 7-31 Functional groups (KEGG Modules) upregulated in the SPF-1 and SPF-2 gut microbial communities

7.6.4 Microbiota-derived gene signatures

7.6.4.1 Workflow

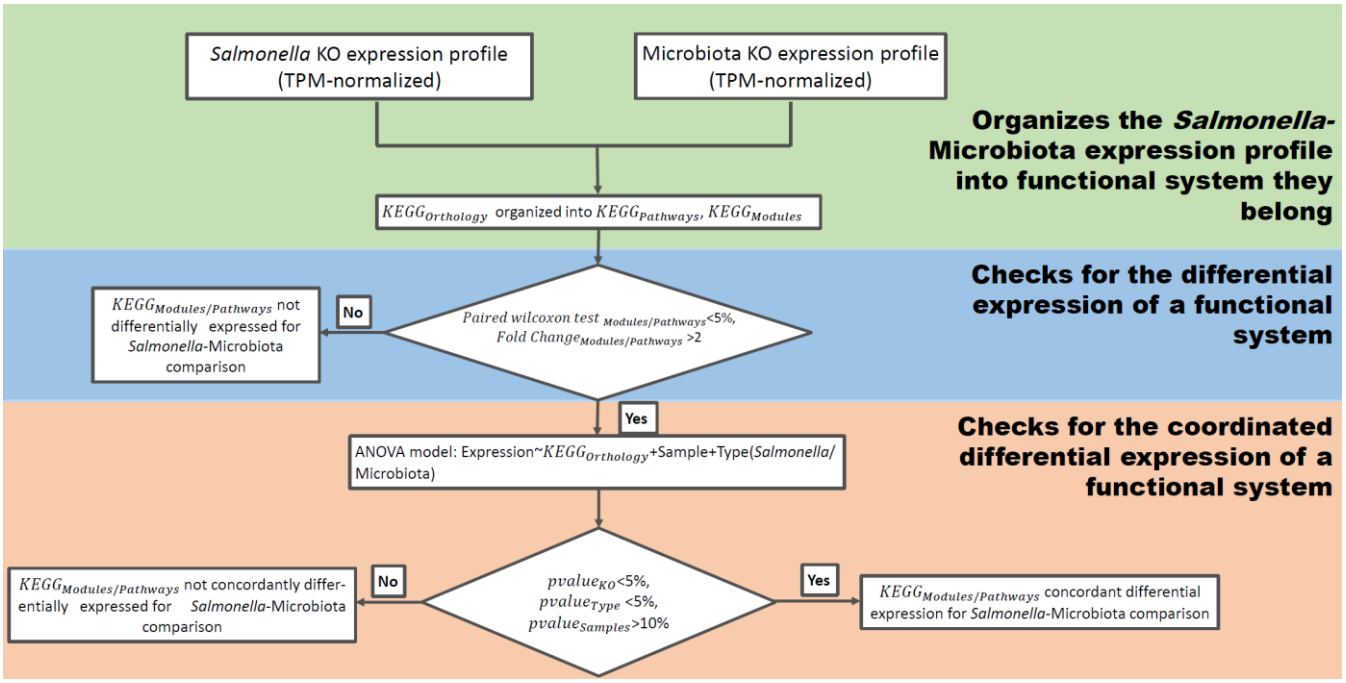


Figure 7-32 Workflow employed to identify potential microbiota-derived gene signatures associated with *S.* Tm infection.

In the previous sections, we performed an associative study by which we identified few potential functional and taxonomic signatures, which could be associated with the *S*. Tm gene signatures in the SPF-1 and SPF-2 conditions. In this section, we move from associative study to identifying the potential functional interactions between *S. Tm* and the gut microbial community. The murine gut microbiota catalogue by Lesker et al. (Manuscript under preparation) not only served as a reference to identify the active members and the functional systems that could be associated with *S*. Tm infection, they also provided the information on the gene repertoire of the active members of the community. We took advantage of this information to infer the potential functional interactions between *S*. Tm and the active members of the SPF-1 and SPF-2 gut microbial community.

We were interested in identifying those functional systems that were active in *S*. Tm, but not in the gut microbiota and vice versa. To obtain such *S.* Tm-microbiota interactions, we employed a methodology as shown above in **Figure 7-32**.

Briefly, we were interested in identifying the functional systems (KEGG Modules or KEGG Pathways) that were active in the microbial community when compared to *S*. Tm in the same gut environment and vice versa. For this,

1. We organized the individual genes in *S.* Tm and the rest of the microbiota into the KEGG Modules/Pathways they belong.
2. We compared if any functional system is highly expressed in either *S.* Tm or the microbiota (Fold change >2 and Wilcoxon test p-value <5%). However, not all the genes constituting such functional systems would show a similar difference in their expression pattern.
3. To enrich for those functional systems that possess most of its constituent genes showing coordinated differential expression, we employed an ANOVA model, which allowed us to pick only those functional systems, whose variance is mainly due to the genes that perform the function and not the variance due to the samples.

We performed this methodology for the expression profiles from both SPF-1 and SPF-2 conditions. The functional systems that were highly expressed in microbiota when compared to *S.* Tm were termed 'microbiota-derived signatures'. The functional systems that were highly expressed in *S.* Tm when compared to microbiota were considered '*S.* Tm *in vivo* fitness signatures'.

SPF-1 gut microbiota environment is more susceptible to *S.* Tm infection in comparison to the SPF-2 gut microbiota environment in cecal tissue. It is essential to identify the functional systems that were active under SPF-1 and SPF-2 conditions in *S.* Tm and the same functional systems' possible activity in the other members of the gut microbiota. From this, we could infer 1). The functional potentials required by the gut microbiota to protect against enteropathogenic infections 2). The activity level of similar functional systems in *S.* Tm and other members of the microbiota in the same gut environment that characterize the *S.* Tm infection *in vivo*.

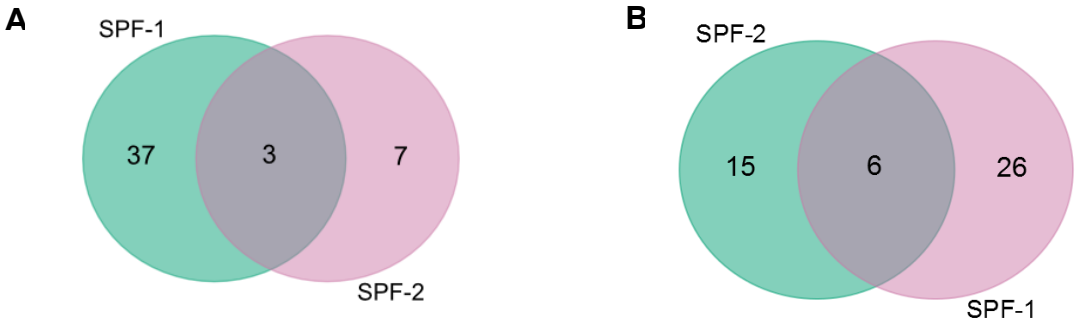


Figure 7-33 Venn diagram representing the Modules (A) and Pathways (B) that were highly expressed in the gut microbiota environment when compared to *S.* Tm.

Amongst modules that were identified as microbiota-derived functional signatures in the SPF-1 and SPF-2 gut microbial communities, there were 3 modules upregulated in microbiota in both SPF-1 and SPF-2 in comparison to *S*. Tm. These include biotin transport system, guanine ribonucleotide biosynthesis and pyrimidine deoxyribonucleotide biosynthesis. Whereas there were 37 (represented by the SPF-1 unique portion in venn diagram **Figure 7-33-A** and represented by the SPF-2 unique portion in venn diagram **Figure 7-33-B**) and 7 unique microbial functional signatures in SPF-1 and SPF-2 gut microbial communities respectively.

Amongst pathways that were identified as microbiota-derived functional signatures common in the SPF-1 and SPF-2 gut microbial communities, there were 6 pathways upregulated in microbiota in both SPF-1 and SPF-2 conditions in comparison to *S*. Tm. These include taurine and hypotaurine metabolism, pyrimidine metabolism, D-glutamine and D-glutamate metabolism, insulin resistance, valine, leucine and isoleucine degradation and beta-lactam resistance. Whereas there were 26 and 15 unique microbial functional signatures in the SPF-1 and SPF-2 conditions respectively. The significance of the functional systems inferred are discussed in detail in the sections below.

7.6.4.2 Microbiota-derived signatures in the SPF-1 microbiota community:

A module M00741 responsible for producing succinyl-CoA from propanoyl-CoA was observed highly expressed in the microbiota when compared to *S*. Tm in SPF-1 condition. A recent study reported that microbiota-derived succinates were essential for *S*. Tm virulence *in vivo* (Spiga et al., 2017). They also observed a complete utilization of TCA cycle by *S*. Tm while causing infection. In our present study, we also observed an upregulation of microbiota-derived succinate from propanoyl-CoA and a complete utilization of TCA Cycle (**Figure 7-21**) in SPF-1 condition, which could be a possible reason for *S*.Tm infection. While the *Clostridiales* from the active members of the community encode three KEGG Orthologs required that constitute this module K01847, K01848 and K01966, *Oscillibacter* could encode only K01847.

Apart from this, many metal-ion transport systems like iron complex, nickel, manganese/zinc, microbiota-derived nutrients like biotin biosynthesis were also upregulated in microbiota. Correspondingly, most of the active members of the SPF-1 microbiota community have the functional potentials of these functional signatures. One of the active members of the SPF-1community belonging to order *Lactobacillales* do not possess most of these active functional systems.

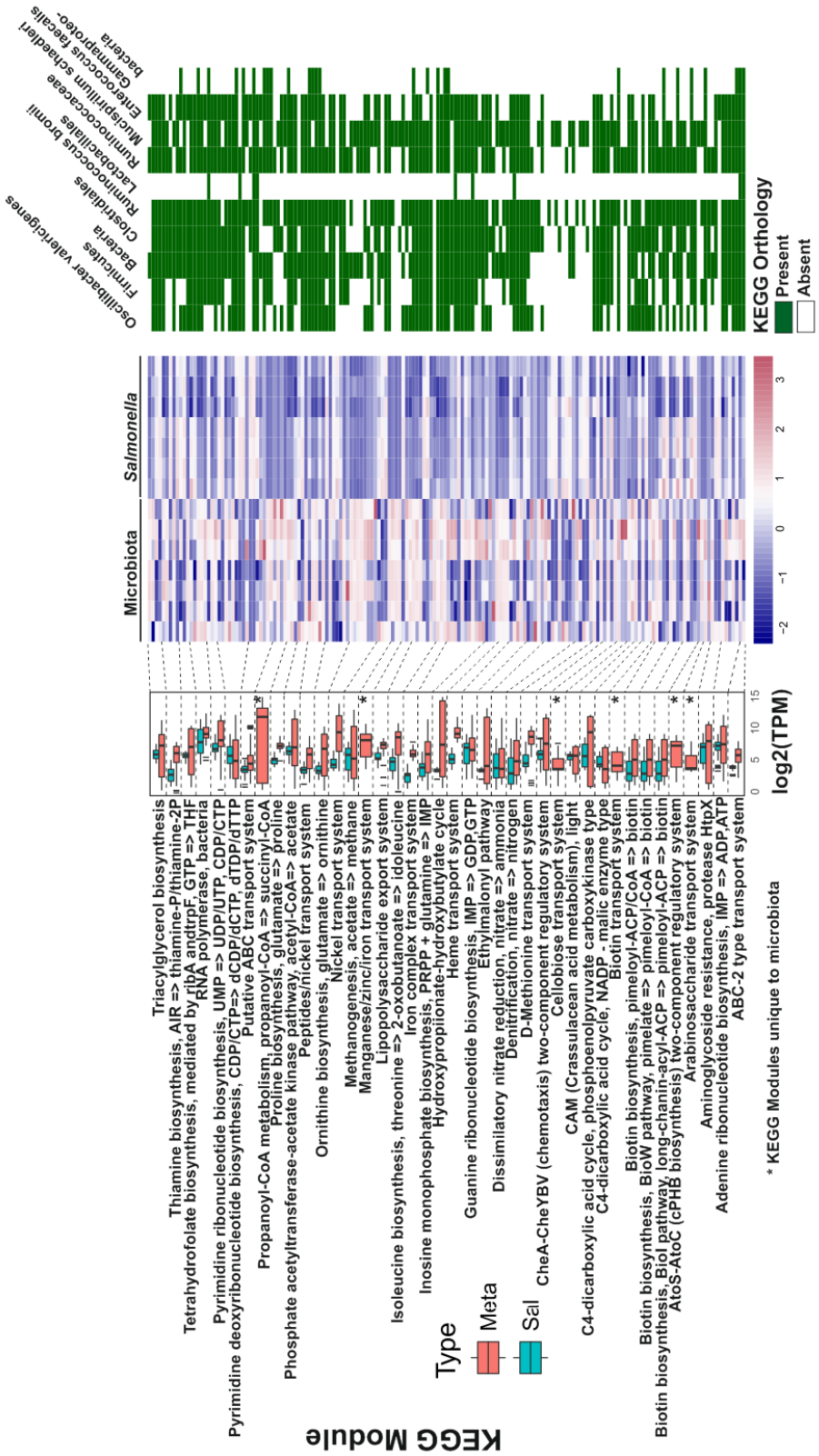


Figure 7-34 SPF-1 microbiota-derived signatures (KEGG Modules) associated with *S.* Tm infection

The boxplot on the left represents the expression levels of the KEGG Modules in both *Salmonella* (blue) and the rest of the gut microbiota (red) in the SPF-1 condition. The functions denoted by asterisk represent the functional systems not inherent in the *S.* Tm genome. The heatmap represents the expression profile of KEGG Orthology belonging to the KEGG Modules highlighted in the boxplot on the left. To the right of this, is the heatmap denoting the presence/absence of these KEGG Orthologies in the genome of the highly active members of the SPF-1 gut microbial community.

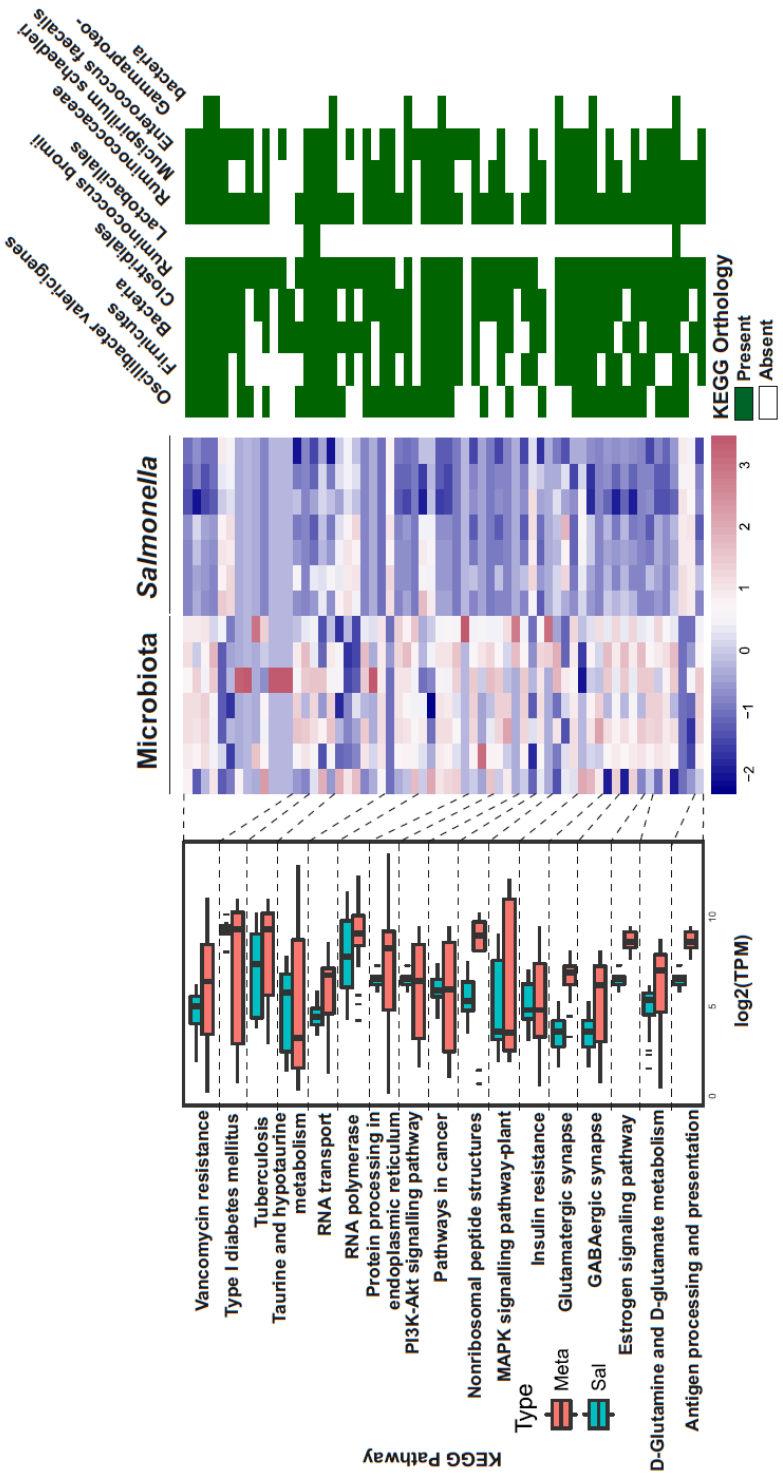


Figure 7-35 SPF-1 microbiota-derived signatures (KEGG Pathways) associated with *S.* Tm infection

The boxplot on the left represents the expression levels of the KEGG Pathways in both *Salmonella* (blue) and the rest of the gut microbiota (red) in the SPF-1 condition. The functions denoted by asterisk represent the functional systems not inherent in the *S.* Tm genome. The heatmap represent the expression profile of KEGG Orthology belonging to the KEGG Pathways highlighted in the boxplot on the left. To the right of this, is the heatmap denoting the presence/absence of these KEGG Orthologies in the genome of the highly active members of the SPF-1 gut microbial community.

Similarly, we observed the KEGG Pathways that were highly expressed in SPF-1 gut microbiota in comparison to *S.* Tm (**Figure 7-34**). The functional systems like

glutamergic synapse, D-glutamate metabolism were highly expressed and most of the highly active members of the SPF-1 community possess the genetic reporteroire to express these functions. All but one active members of the SPF-1 microbiota community (belonging to order *Lactobacillales*) do not possess most of these functional reportoires. It is interesting to note that the common inhabitants of the gut environment like *Ruminococcus bromii* and *Enterococcus faecelis* also possess these functional potentials that could be associated with *S*. Tm infection. Also, we could observe a redundancy in functional potentials amongst the active members of the SPF-1 community that could aid *S*. Tm to adapt, survive and cause infection. This might be indicating that instead of one "keystone species" to provide protection against *S*. Tm, the microbial ecosystem could encompass many active members possessing functional potentials able to provide protection against *in vivo* fitness for *S*. Tm.

7.6.4.3 Microbiata-derived signatures in the SPF-2 gut microbial community:

A study published by *Mardinoglu* et al., (Mardinoglu et al., 2015) reported that the gut microbiota regulates glutathione metabolism in conventionally raised mouse models. Glutathione is one of the key anti-oxidants whose deficiency lead to complex disorders. Here, we observed a module M00118 responsible for glutathione metabolism, with two KEGG Orthologs K01919 and K01920, were highly expressed in SPF-2 microbial community in comparison to *S*. Tm. In addition, a module M00582 with 8 KEGG Orthologs responsible for Energy coupling-factor transport, Vitamin-B12 transport were upregulated in SPF-2 microbiota. Correspondingly, amongst the active members of the community, *Lachnospiraceae* possess the functional potentials for most of the KOs from such functional signatures. This is shown in **Figure 7-36**. This might suggest a significant role of *Lachnospiraceae* to provide microbiota-derived compounds essential for protection against enteropathogenic infections. In addition, it is interesting to note the microbe of order *Clostridiales* and *Lachnospiraceae bacterium A4* possess a unique module M00581 responsible for biotin transport highly expressed in the SPF-2 condition.

In spite of the differences in their taxonomic lineages, most of the active members of the SPF-2 microbial community were capable of expressing many of the microbiota-derived functional signatures. This emphasizes the significance of the presence of multiple microbes possessing redundant functions to provide protection against pathogens, which was also observed in (Kang et al., 2015).

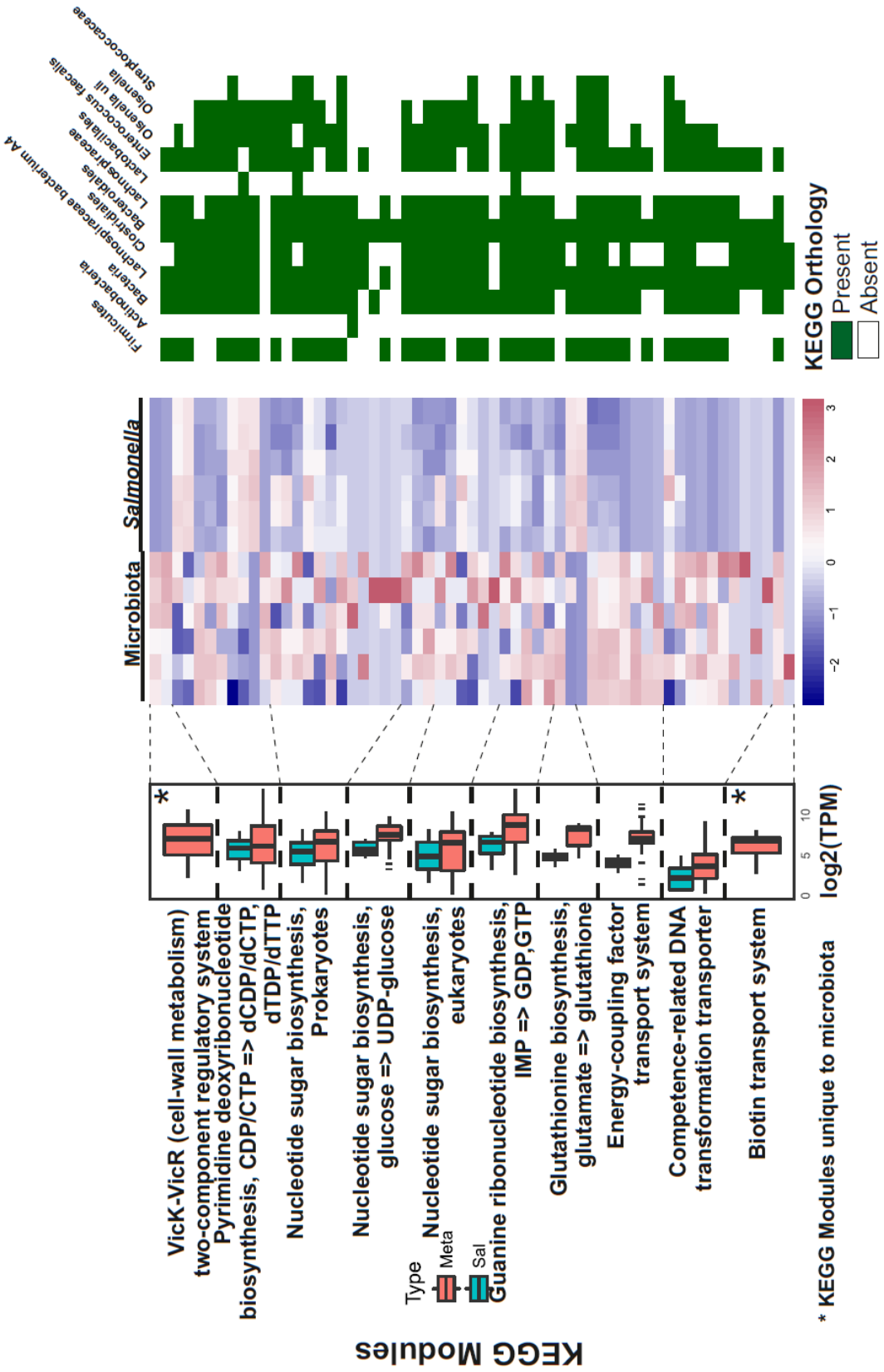


Figure 7-36 Microbiota-derived signatures (KEGG Modules) in the SPF-2 gut

The boxplot on the left represent the expression levels of the KEGG Modules in both *Salmonella* (blue) and the rest of the gut microbiota (red) in the SPF-2 condition. The functions denoted by asterisk represent the functional systems not inherent in the *S.* Tm genome. The heatmap represents the expression profile of KEGG Orthology belonging to the KEGG Modules highlighted in the boxplot on the left. The dotted line represent the KEGG Orthologies belonging to the KEGG Modules. To the right of this, is the heatmap denoting the presence/absence of these KEGG Orthologies in the genome of the highly active members of the SPF-2 gut microbial community.

microbiota community associated with *S*. Tm infection

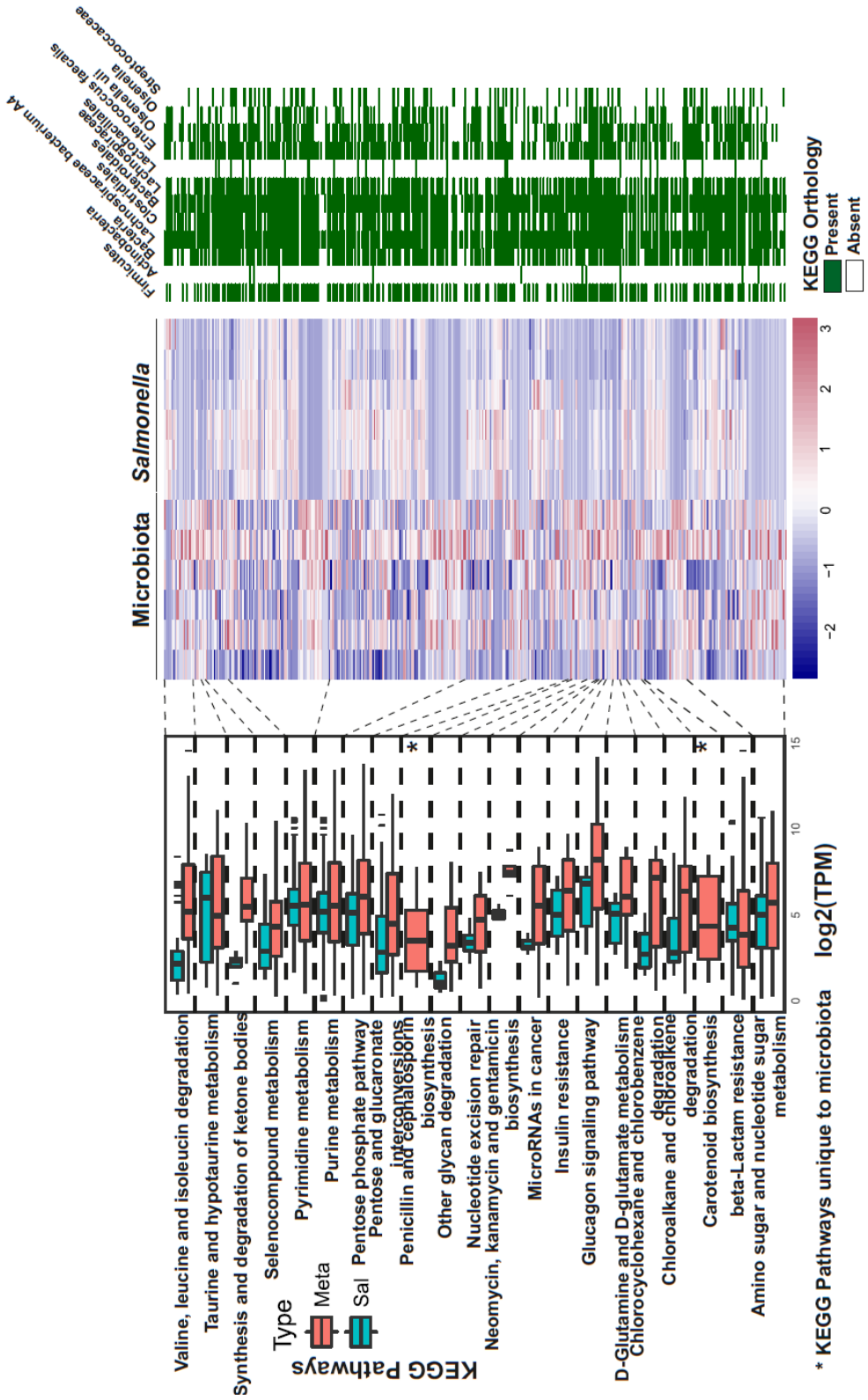


Figure 7-37 Microbiota-derived signatures (KEGG Pathways) in the SPF-2 gut

The boxplot on the left represents the expression levels of the KEGG Pathways in both *Salmonella* (blue) and the rest of the gut microbiota (red) in the SPF-2 condition. The functions denoted by asterisk represent the functional systems not inherent in the *S*. Tm genome. The heatmap represent the the expression profile of KEGG Orthology belonging to the KEGG Pathways highlighted in the boxplot on the left. The dotted line represent the KEGG Orthologies belonging to the KEGG Pathways. To the right of this, is the heatmap denoting the presence/absence of these KEGG orthologies in the genome of the highly active members of the SPF-2 gut microbial community.

microbiota community associated with S. Tm infection

7.6.4.4 *S.* Tm *in vivo* fitness:

In the previous section, we observed those microbiota-derived functional signatures that could be associated with the *S.* Tm virulence. In this section, we would look at the functional signatures in *S.* Tm that were highly expressed when compared to the gut microbial community and could be associated with its virulence. For this, we picked those KEGG Modules and KEGG Pathways that were highly expressed in *S.* Tm when compared to other members of the gut microbial community.

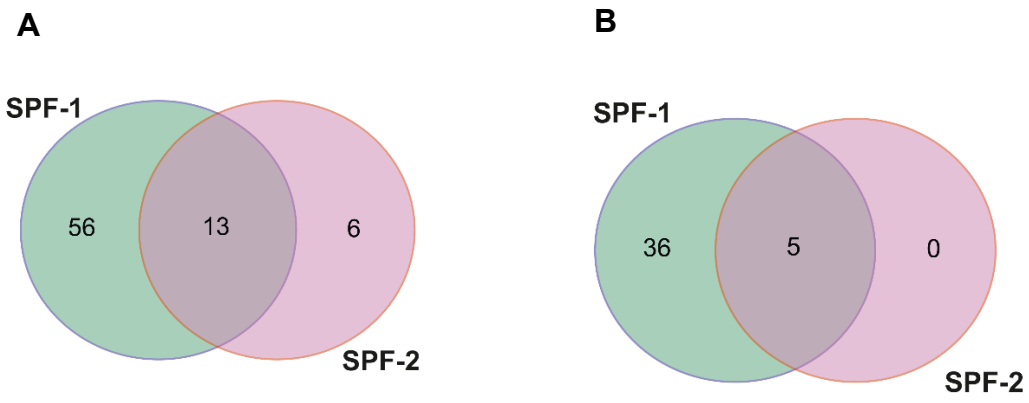


Figure 7-38 Venn diagram representing the KEGG Modules (A) and KEGG Pathways (B) highly expressed in *Salmonella* when compared to microbiota.

Venn diagram **Figure 7-38** shows the number of modules that were upregulated in *S.* Tm uniquely and in both the gut microbiota environments. We discuss in detail regarding the unique functional systems (KEGG Modules and KEGG Pathways) upregulated in *S.* Tm that could be attributed to its infection potential. Alongside these KEGG Modules, we observed the active members of the other gut microbiota environment to see whether those microbes would have the functional potentials to resist/aid in *S.* Tm infection.

7.6.4.5 KEGG Modules uniquely upregulated in *Salmonella* in comparison to the SPF-1 gut microbial community:

We picked the 56 KEGG Modules that were upregulated in *S.* Tm in the SPF-1 environment alone (represented by the SPF-1 unique portion in venn diagram **Figure 7-38-A**). Alongside the functional signatures, we observed the active and distinct members of the SPF-2 microbiota community to get an insight on the potential protection that could be provided by these microbes had they been in the SPF-1 community.

In the SPF-1 condition which is more susceptible to *S*. Tm infection in cecal tissue, *S*. Tm upregulated the uptake systems for the simple sugar nutrients, in comparison to SPF-1 microbial community like tricarboxylic acids, C4-dicarboxylic acids, glycerolphosphate, fructose, glucose, alpha and beta-glucoside. One of the active members of the SPF-2 community belonging to order *Lachnospiraceae* possess similar functional potentials to uptake these nutrients. This observation suggests that, had these microbes were present in SPF-1 environments; they could have provided competition against *S*. Tm to feed on these simple sugars, thereby reducing the

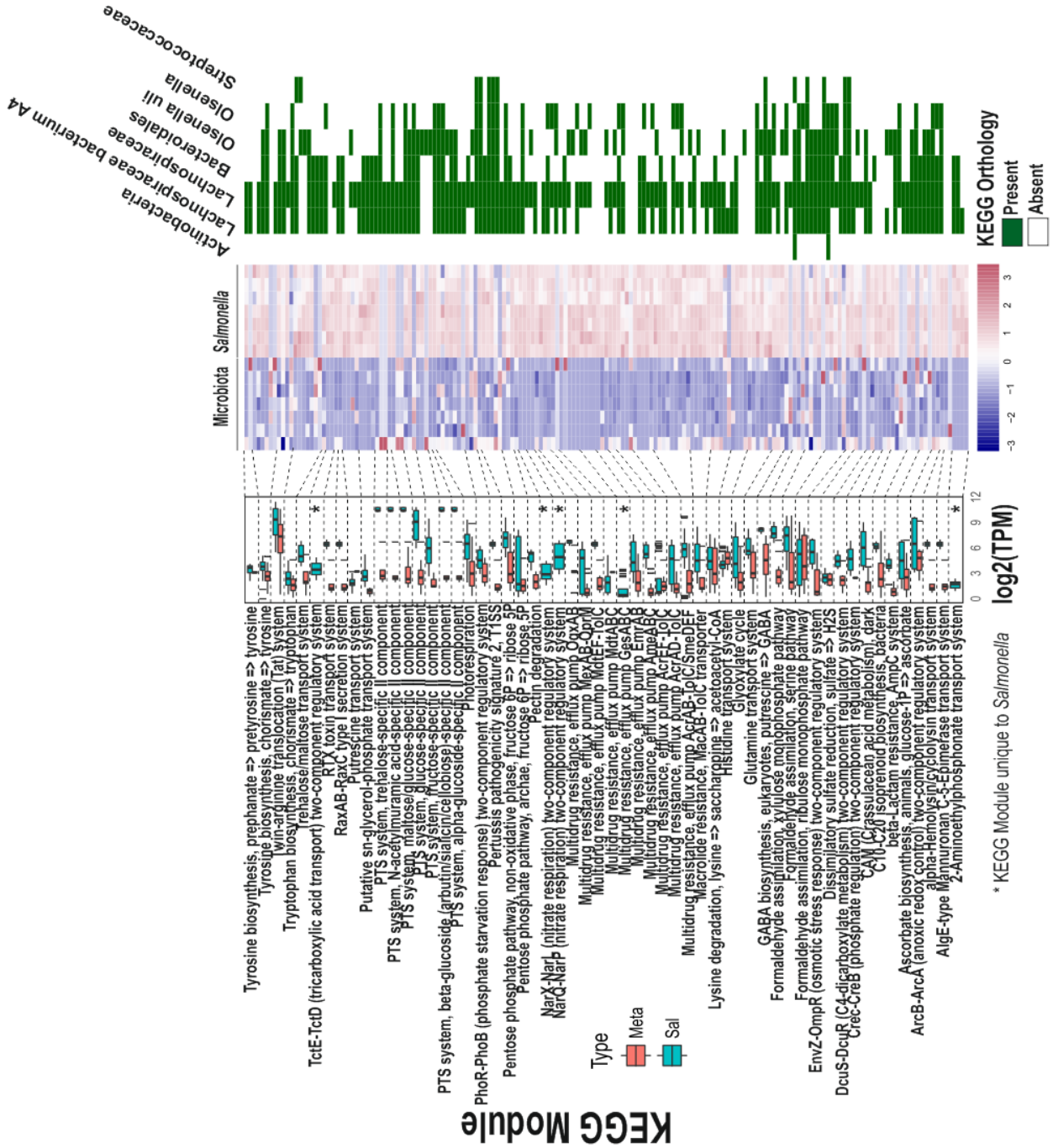


Figure 7-39 *S*. Tm signatures (KEGG Modules) in SPF-1 gut microbial community associated with its *in vivo* fitness

The boxplot on the left represent the expression levels of the KEGG Modules in both *Salmonella* (blue) and the rest of the gut microbiota (red) in the SPF-1 condition. The functions denoted by asterisk represent the functional systems inherent in the *S*. Tm genome. The heatmap represents the expression profile of KEGG Orthology belonging to the KEGG Modules highlighted in the boxplot on the left. To the right of this, is the heatmap denoting the presence/absence of these KEGG Orthologies in the genome of the highly active members of the SPF-1 gut microbial community.

chances of *S*. Tm infection in SPF-1 environment in cecal tissue and reducing the chances of systemic infection.

We also observed *S*. Tm upregulating RTX toxin system and Type-I secretion systems in SPF-1 environment. An inherent functional system belonging to *S*. Tm responsible for nitrate respiration were also upregulated, which possibly provide *S*. Tm an edge over other microbes for survival and thereby causing infection.

7.6.4.6 KEGG Pathways uniquely upregulated in *Salmonella* in comparison to the SPF-1 gut microbial community:

Similar to the above section, we picked the 36 KEGG Pathways that were uniquely upregulated in *S*. Tm in comparison to the SPF-1 gut microbiota environment (represented by the SPF-1 unique portion in venn diagram **Figure 7-38-B**). Amongst them, the genes involved in degradation of multiple nutrient sources like linoleic acid, glycosaminoglycans and glutathione metabolism were upregulated in *S*. Tm. This is depicted by **Figure 7-40**.

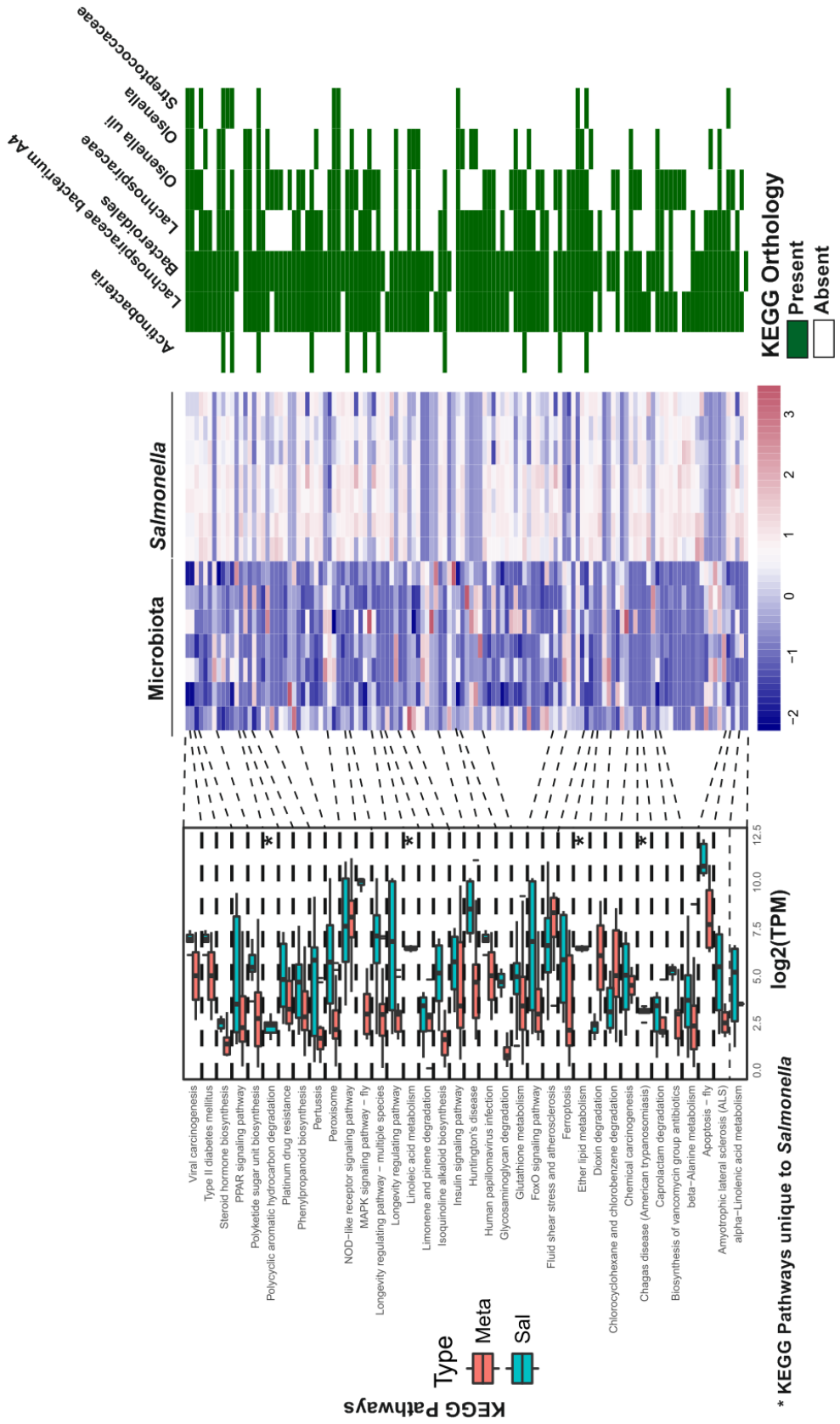


Figure 7-40 *S.* Tm signatures (KEGG Pathways) in SPF-1 gut microbial community associated with its *in vivo* fitness

The boxplot on the left represents the expression levels of the KEGG Pathways in both *Salmonella* (blue) and the rest of the gut microbiota (red) in the SPF-1 condition. The functions denoted by asterisk represent the functional systems inherent in the *S*. Tm genome. The heatmap represents the expression profile of KEGG Orthology belonging to the KEGG Pathways highlighted in the boxplot on the left. To the right of this, is the heatmap denoting the presence/absence of these KEGG Orthologies in the genome of the highly active members of the SPF-1 gut microbial community.

7.6.4.7 KEGG Modules uniquely upregulated in *Salmonella* in comparison to the SPF-2 gut microbial community:

Similar to previous section, in this section we identify those functional systems that were upregulated uniquely in *S*. Tm in comparison to other microbes in the SPF-2 condition. As shown in **Figure 7-41**, there were 6 modules upregulated uniquely in *S*. Tm in SPF-2 environment. Amongst them, three of these functions were not part of the functional repertoire of either SPF-1 or SPF-2 environment. Notably, A KEGG Module M005422 encoding a non-flagellar Type-III Secretion System (T3SS), which is inherent only to *S*. Tm genome, was highly expressed in *S*. Tm in the SPF-2 environment and no other microbes in both SPF-1 and SPF-2 microbial ecosystem possessed this functional potential. Based on this, one could also hypothesis that this non-flagellar T3SS system could be a key functional signature that provides an edge for *S*. Tm to cause infection. **Figure 7-41** depict this comparison.

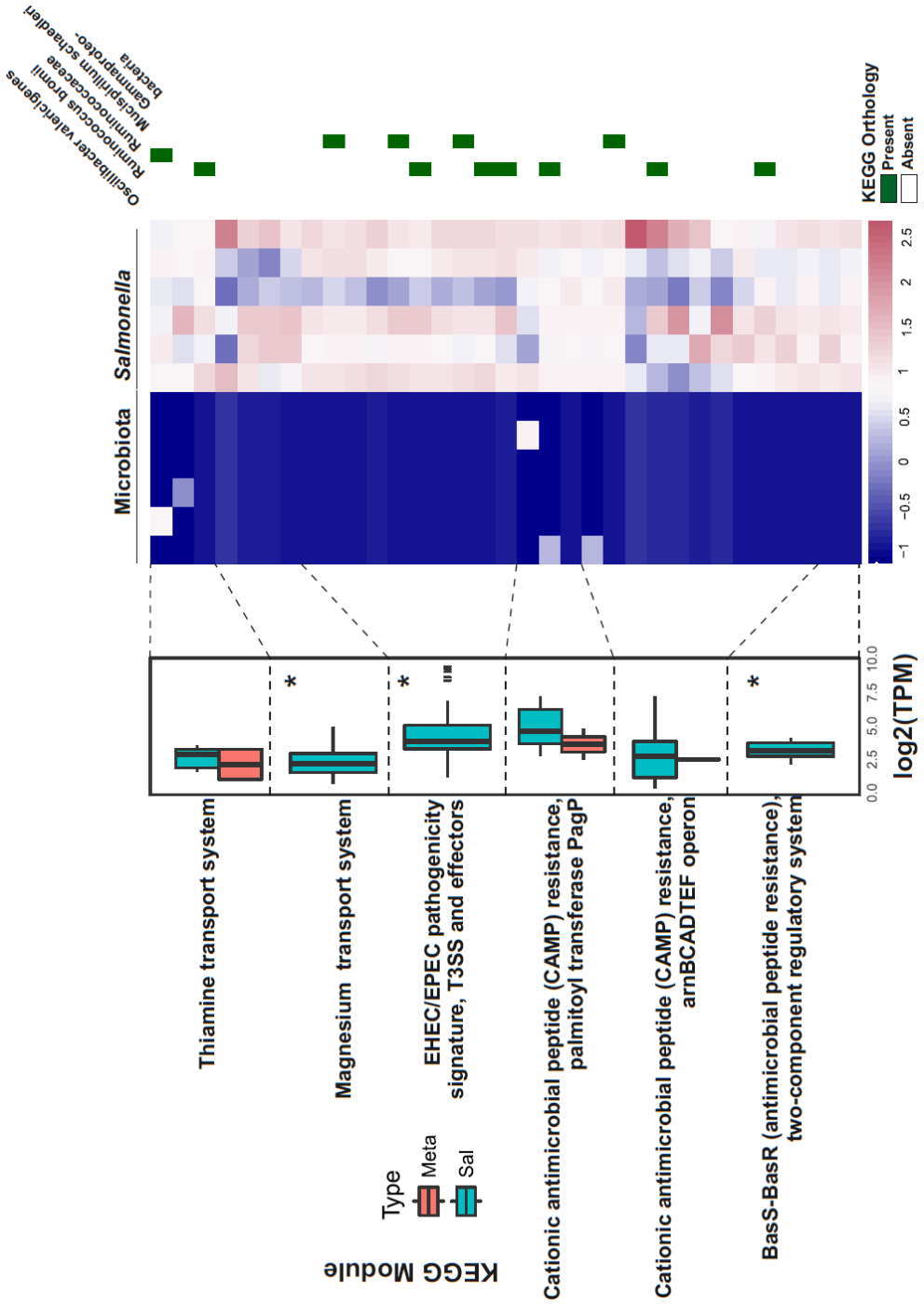


Figure 7-41 S. Tm microbiota-derived signatures (KEGG Pathways) in SPF-2 gut microbial community associated with its *in vivo* fitness

The boxplot on the left represent the expression levels of the KEGG Pathways in both *Salmonella* (blue) and the rest of the gut microbiota (red) in the SPF-2 condition. The functions denoted by asterisk represent the functional systems inherent in the *S.* Tm genome. The heatmap represents the expression profile of KEGG Orthology belonging to the KEGG Pathways highlighted in the boxplot on the left. To the right of this, is the heatmap denoting the presence/absence of these KEGG Orthologies in the genome of the highly active members of the SPF-2 gut microbial community.

We identified no KEGG Pathways uniquely highly expressed in S. Tm when compared to SPF-2 gut microbial community.

7.6.4.8 KEGG Modules upregulated in *S*. Tm in comparison to both the SPF-1 and SPF-2 gut microbial community:

We identified 13 KEGG Modules upregulated in S. Tm in comparison to both the SPF-1 and SPF-2 gut microbial community (represented by the intersection portion of the venn diagram **Figure 7-38-A**). These KEGG Modules includes multidrug resistance efflux pumps and *S.* Tm inherent two-component systems responsible for anaerobic respiration- TorS/TorR system (Paiva et al., 2009), envelope stress response (BasS-BaeR) and virulence- SsrA/SsrB (Choi et al., 2010), PhoQ/PhoP (Tu et al., 2006), BarA/UvrY (Palaniyandi et al., 2012) . Correspondingly, one of the active members of the SPF-2 community belonging to family *Lachnospiraceae* possess the gene reportoires to express and thereby counter *S*. Tm virulence. This is evident in the **Figure 7-42**.

However, in the SPF-1 gut microbial community, which is more susceptible to *S*. Tm infection, the common *S*. Tm *in vivo* signatures were expressed higher than the other members of the gut microbial community. This is mainly because, most of the active members of the community do not possess gene repertoire to encode them to compete against *S.* Tm in the SPF-1 gut microbial community. We can also observe that *S*. Tm express these signatures higher in SPF-1 community in comparison to that of the SPF-2 community (evident from the boxplots at the left end of the figures **Figure 7-42**, **Figure 7-43**). This again highlights the significance of the presence of the active members of the gut microbiota in protection against enteropathogenic infections.

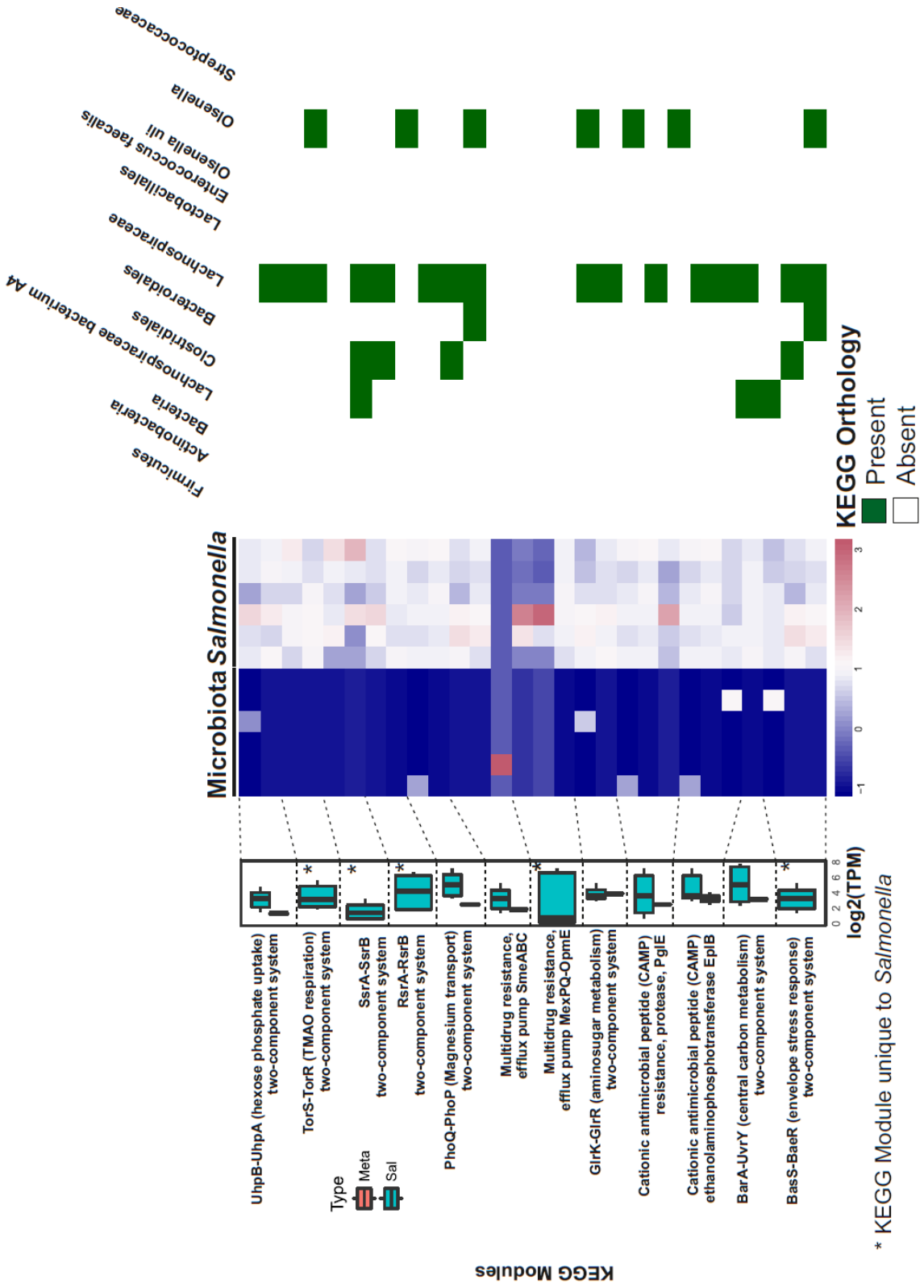


Figure 7-42 *S*. Tm microbiota-derived signatures (KEGG Modules) common in both the SPF-1 and SPF-2 gut microbial community associated with its *in vivo* fitness

The boxplot on the left represent the expression levels of the KEGG Modules in both the *S*. Tm (blue) and the rest of the gut microbiota (red) in the SPF-2 condition. The functions denoted by asterisk represent the functional systems inherent in the *S*. Tm genome. The heatmap represents the expression profile of KEGG Orthology belonging to the KEGG Modules highlighted in the boxplot on the left. To the right of this, is the heatmap denoting the presence/absence of these KEGG Orthologies in the genome of the highly active members of the SPF-2 gut microbial community.

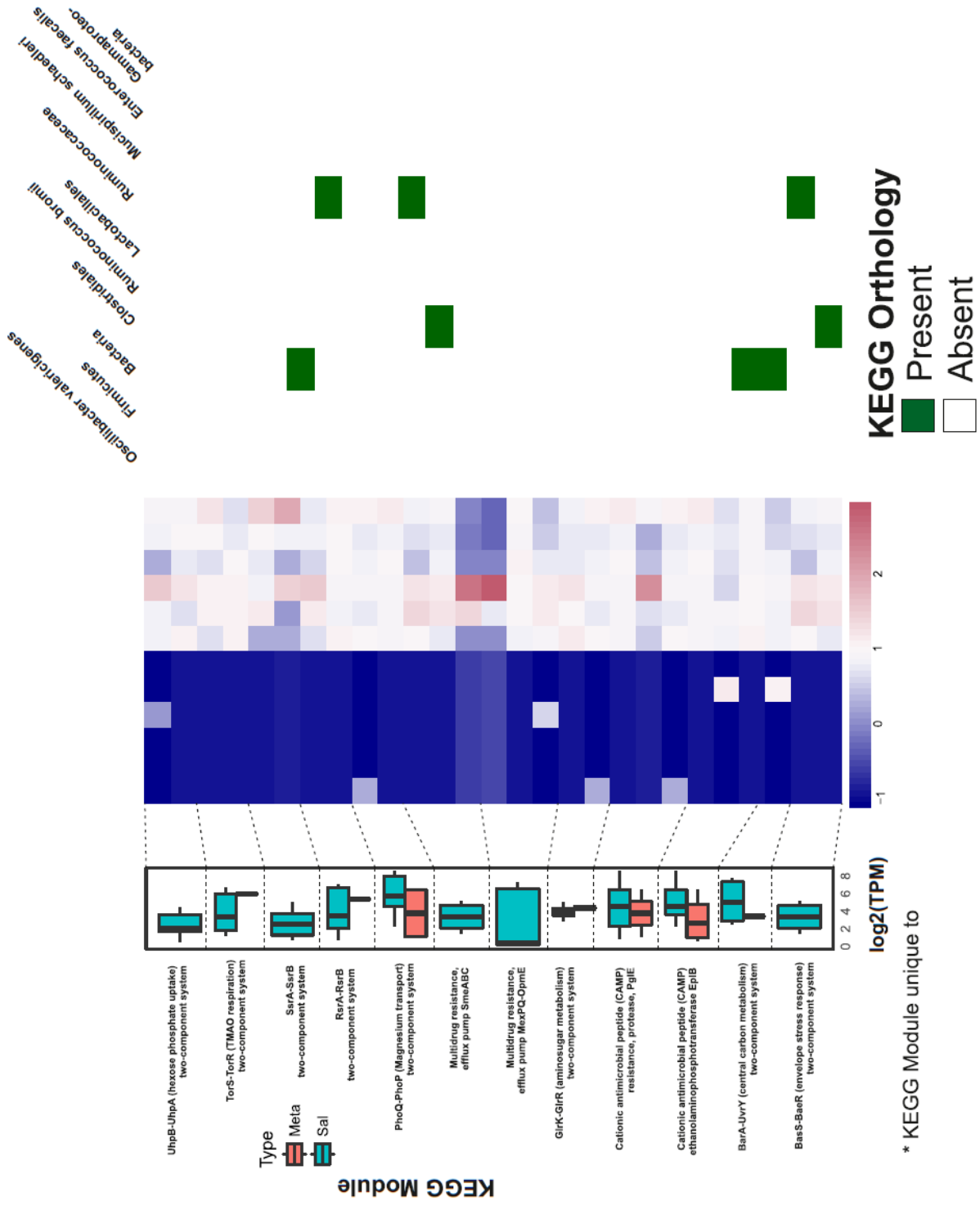


Figure 7-43 *S*. Tm microbiota-derived signatures (KEGG Modules) common in both the SPF-1 and SPF-2 gut microbial community associated with its *in vivo* fitness

The boxplot on the left represent the expression levels of the KEGG Modules in both the *S*. Tm (blue) and the rest of the gut microbiota (red) in the SPF-1 condition. The functions denoted by asterisk represent the functional systems inherent in the *S*. Tm genome. The heatmap represents the expression profile of KEGG Orthology belonging to the KEGG Modules highlighted in the boxplot on the left. To the right of this, is the heatmap denoting the presence/absence of these KEGG Orthologies in the genome of the highly active members of the SPF-1 gut microbial community.

To summarize this section, *S*. Tm has a robust transcriptional machinery, which causes infection exploiting the microbiota-derived nutrients under dysbiotic gut microbial

ecosystems. In spite of a drastic change in the microbial composition and functional potentials of the gut microbial ecosystems under study, *S*. Tm rewires few specific transcriptional machineries in response to the microenvironments created by the combinatorial effect of the host and the gut microbial ecosystem, and cause infecton in cecal content. *S*. Tm expression profile in cecal content could be an indicator of its virulence in the cecal tissue since initiation and maintenance of few transcriptional machineries could be observed in its expression profile. On the hindsight, the gut microbial community should be constituted by many microbes possessing functional potentials that could be associated with protection. This emphasis the hypothesis that a gut microbial environment with functional redundancy would provide better protection against enteropathogenic infections (redundancy hypothesis), as opposed to one microbial signature providing protection (keystone species).

7.7 Increase in efficiency of identifying potential *Salmonella*-Microbiota interactions

The *S*. Tm-microbiota functional interactions observed in the previous section, is feasible only due to the availability of a host-specific murine gene catalogues with improved taxonomic and functional resolution (Lesker et al.). A host-specific reference gene catalogue is high essential to identify the active members of a microbial community and the genetic elements highly expressed in such active members. The currently available murine gut microbial gene catalogue (Xiao et al., 2015) provides taxonomic and functional annotations only at a broad level. An improved taxonomic and functional annotations for the metatranscriptomic study was provided by a murine gut microbiome catalogue recently developed by Lesker et al., To compare the efficiency of the new murine gut microbiome catalogue in providing taxonomic annotation with better resolution, we compared the taxonomic classification by comparing the conversion rates to the next higher taxonomy level. For the samples from both SPF-1 and SPF-2 conditions, we observed a better conversion rate of taxonomic annotations from phylum to class and genus to species using the new reference gene catalogue when compared to the one organized by (Xiao et al., 2015) for murine gut microbiota (**Figure 7-44**). If

$$Taxonomy = \{Kingdom, Phylum, Class, Order, Family, Genus, Species\}$$

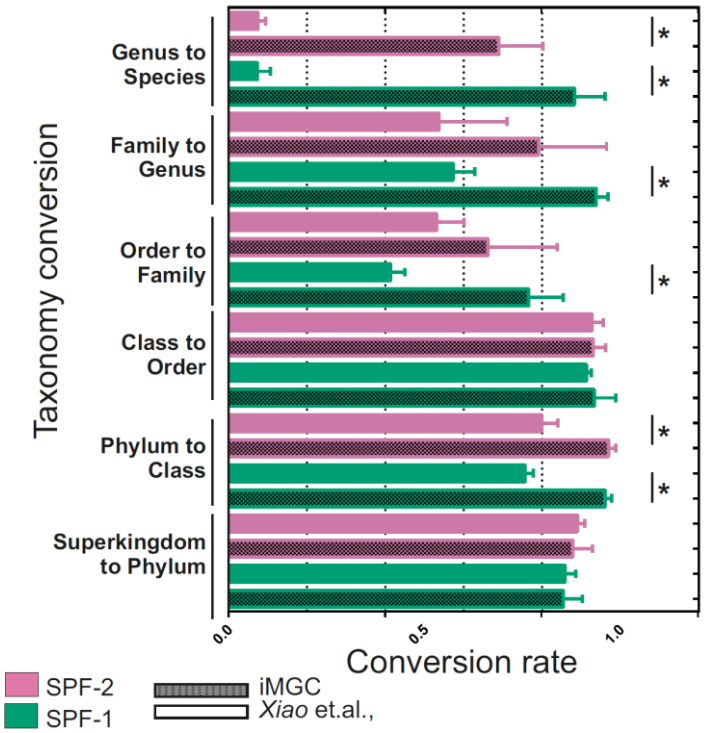
$$Taxonomy\ conversion\ from\ Taxonomy_N\ to\ Taxonomy_{N-1} = {\sum_{i=1}^{N-1} Reads_i}\Big/{Reads_N}$$

Where $Taxonomy_N$ is the higher taxonomy level, $Taxonomy_{N-1}$is the immediate lower taxonomy level and $Reads_i$ represent the number of reads that obtained a taxonomy annotation of $Taxonomy_i$. In order to further improve the ability of the novel murine gut

microbiome reference catalogue to provide improved taxonomic resolution, we employed a novel methodology in the next section.

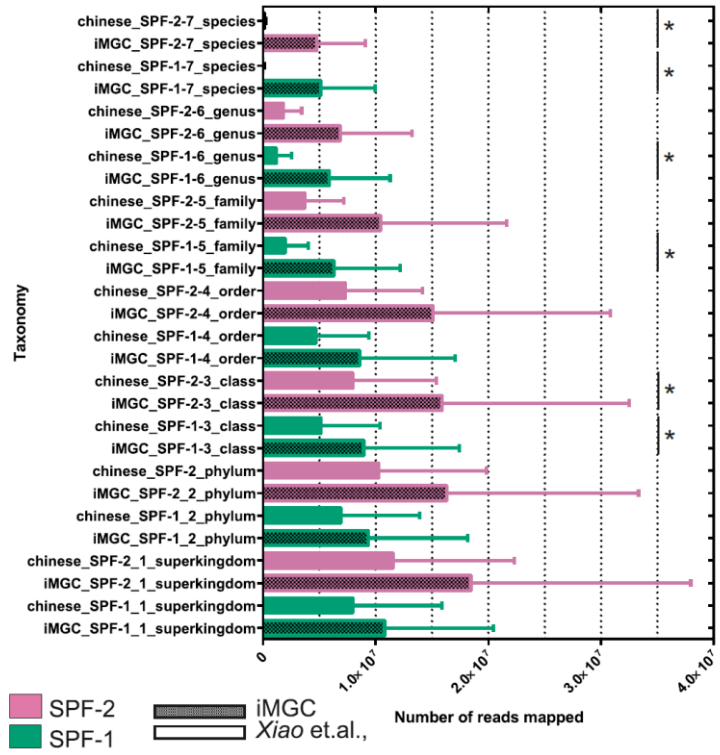
7.7.1 Necessity of mouse-gut specific reference catalogue

Metagenomics estimate the abundance of the members of the microbial communities by assembling the reads into contigs to predict the genes they belong and clustering the contigs into “bins” which potentially represent one member of the microbial community. The bins were assigned taxonomic annotations based on the presence and abundance of gene sets characteristic of previously studied microbe. Though this method works good for well-studied microbes, it is unable to identify the unstudied and unsequenced microbes. 16S rRNA genes could provide better taxonomic assignments in such scenarios, however the 16S rRNA reads were very sparse in the bins to provide better taxonomic assignments and classical assembly approaches cannot completely resolve these issues.



The total number of reads mapped to each taxonomy level from comparing the meta-transcriptome data against the two reference catalogues were obtained. The fraction of reads mapped to all the immediate lower taxonomy levels in comparison to the higher taxonomy levels was plotted. The difference in pattern of shade for the bars indicate the difference in gut catalogue used for the comparative study. The difference in color of the bars indicate the different gut microbial community. The asterisk indicate the statistical significance (Wilcoxon rank-sum test) for the conversion rate between the two gut reference catalogues.

Figure 7-44 Improvement in taxonomic resolution for the SPF-1 and SPF-2 microbial communities using iMGC and Xiao et.al., 2015



The barplot represent the number of reads mapped to a gene with particular taxonomy level annotation, obtained from either of the two murine gut microbiota reference catalogues, with the error bar denoting one std.deviation. The asterisks denote the statistical significance in difference between the number of reads mapped by Wilcoxon rang-sum test.

Figure 7-45 Taxonomic resolution for the SPF-1 and SPF-2 microbial communities using iMGC and Xiao et.al., 2015

Also, we observed the new murine gut microbe reference catalogue was able to provide most of the reads mapped in **Figure 7-44**, with improved taxonomic resolution when compared to that of the earlier version (Xiao et al., 2015) as depicted in **Figure 7-45**.

7.7.2 Integrating MGS Bins to 16S rRNA genes

7.7.2.1 Motivation

Identifying the members of gut microbial communities and inferring their functional potential and their taxonomic lineage is crucial in understanding the role of members within a gut microbial community. Metagenomics provides an opportunity to study and characterize microbes without a need to culture them. It provides a computational framework to annotate the short sequences to the functional units and the taxonomic units they represent.

Metagenomics estimate the abundance of members of the microbial communities by assembling the reads into contigs to predict the genes they encode and clustering the contigs into the “Meta Genome Species bins” (MGS Bins in short), which potentially represent one member of the microbial community. The MGS Bins were assigned taxonomic annotations based on the presence and abundance of gene sets characteristic (or sequence similarity) of previously studied microbes (Thompson et al., 2013). The gene catalogues obtained thereby, with high-resolution taxonomic and functional annotations were essential to enhance the characterization of the murine

microbial communities with a large proportion of uncultured microbes. Though this method works well for well-studied microbes, it is unable to identify the high resolution taxonomy for the not-so-well studied microbes. 16S rRNA genes could provide better taxonomic assignments in such scenarios, however the 16S rRNA gene sequences were very sparse in the bins to provide better taxonomic assignments and classical assembly approaches cannot completely resolve these issues.

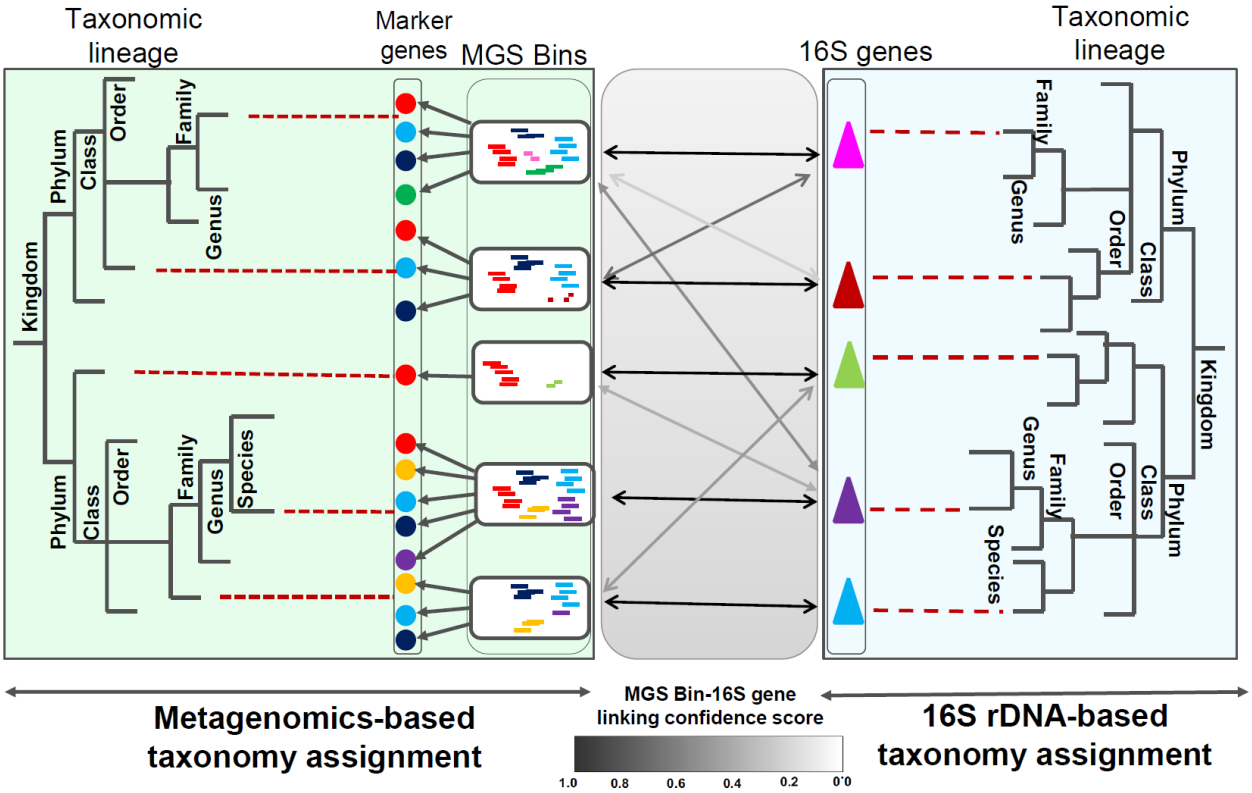


Figure 7-46 Motivation for the MGS bins-16S rRNA gene linking confidence score.

In order to provide an enriched high-resolution taxonomic assignment for the MGS Bin, we employed a novel methodology in integrating the abundance of reconstructed 16S rRNA genes in each MGS Bins to their co-abundance information (obtained from Lesker et al.) using a novel scoring scheme to improve the taxonomic annotation assigned to the MGS Bins.

Recent advances have enabled the efficient reconstruction of 16S rRNA gene sequences from complex metagenome data. Using RAMBL 16S rRNA gene sequences (Zeng et al., 2017) was reconstructed by pooling the short sequences from all the samples by Lesker et al. To this end, we developed an integrated score combining mapping- and correlation-based associations to assign each MGS and 16S rRNA gene sequence. Briefly, all contigs containing reconstructed 16S sequences were identified via BlastN similar to (Mikheenko et al., 2016) by Lesker et al. In parallel, the read-pairs in which one read mapped to a reconstructed 16S sequence and the

other to a contig belonging to an MGS bin was searched by Lesker et al. Finally, all libraries were remapped to all the bins and the 16S rRNA gene sequences to determine their abundances across all samples by Lesker et al. Later, this data was used to estimate correlations between bins and 16S sequences using an abundance co-variance strategy. These individual information were finally integrated using a novel framework to assign the reconstructed 16S rRNA genes to bins. These associations were individually evaluated for taxonomic overlap for bins for which genomes and 16S rRNA genes were a priori known, supporting the validity of this approach. This workflow is depicted in **Figure 7-47**.

7.7.2.2 Workflow

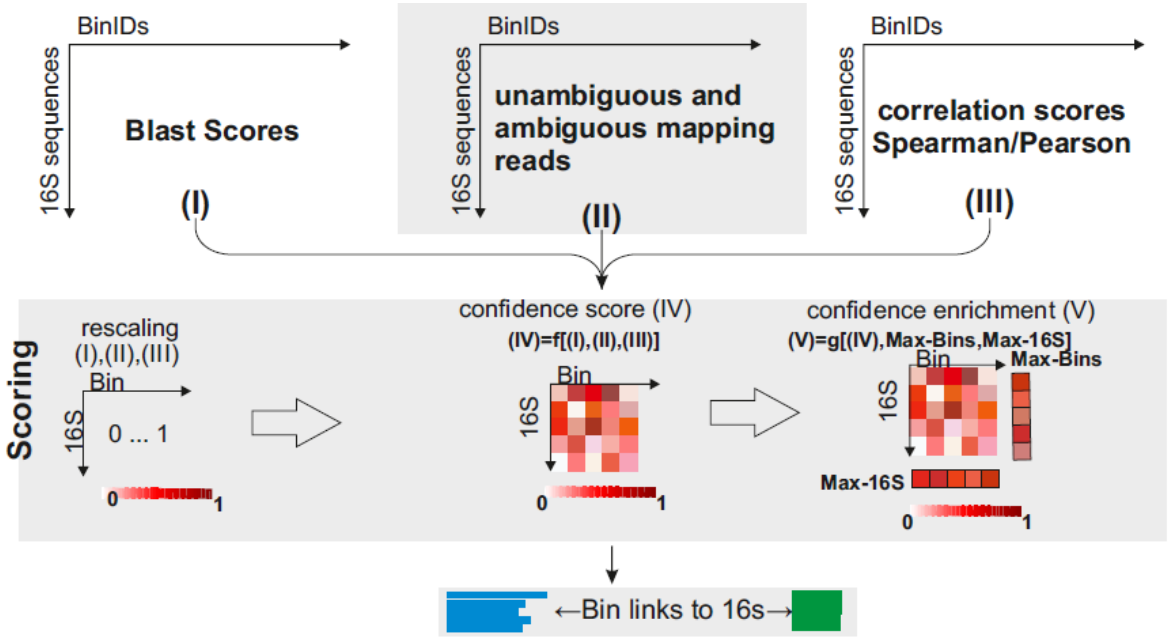


Figure 7-47 Workflow employed to link the 16S rRNA genes to the MGS Bins in the novel murine gut microbiome reference catalogue

7.7.2.2.1 Indirect association (abundance co-variance):

We used the normalized abundance values of the MGS Bins and 16S data to obtain their corresponding correlation (both Pearson and Spearman). We assumed that the correlation (Spearman correlation and Pearson correlation) between a MGS Bin abundance and its corresponding 16S abundance (a true positive direct relationship) should be strongly positively correlated; though need not necessarily be the top-most. In order to enrich such relationships, we unified these correlation values between the MGS Bins and 16S rRNA genes by taking geometric mean of both the correlation values between each MGS Bin and 16S rRNA gene and assigning a negative sign if either of these correlation values is negative.

$$V(x,y) = value[I(x,y)] = \sqrt{abs(P(x,y)) * abs(S(x,y))}$$

$$Sg(x,y) = sign[(I(x,y))] = \begin{cases} -, any\ [P(x,y), S(x,y)] < 0 \\ +, \quad else \end{cases}$$

$$I(x,y) = V(x,y) * Sg(x,y)$$

Where

$P(x,y)$= Pearson correlation between a 16S rRNA gene 'x' and MGS Bin 'y'

$S(x,y)$= Spearman correlation between a 16S rRNA gene 'x' and MGS Bin 'y'

$I(x,y)$= Integrated correlation between a 16S rRNA gene 'x' and MGS Bin 'y'

7.7.2.2.2 Direct association:

1. Mapping MGS Bins to 16S rRNA genes [$M(x,y)$]: These quantify the fraction of reads in a MGS Bin 'y' containing reads in 16S rRNA gene 'x' by mapping the reads in MGS Bin 'y' to the 16S rRNA gene 'x'. We normalized the number of uniquely mapped reads in MGS Bin 'y' to a 16S rRNA gene 'x' by the total number of 16S reads mapped to the MGS Bin 'y'.

$$M\ (x,y)\ = \frac{M(x,y)}{\sum_{i=1}^{n} M(x,i)}$$

2. BLAST MGS Bins to 16S rRNA genes [$B(x,y)$]: These quantify the fraction of reads in a MGS Bin 'y' containing reads in 16S rRNA gene 'x' by finding alignment of the reads in 16S rRNA gene 'x' to the MGS Bin 'y' using BLAST. We normalized the number of uniquely mapped reads in MGS Bin 'y' to a 16S rRNA gene 'x' by the maximum of reads from 16S rRNA genes mapped to the MGS Bin 'y'.

$$B\ (x,y)\ = \frac{B(x,y)}{max_{0<y\leq n}\ B(x,y)}$$

7.7.2.2.3 Integrating the Direct and Indirect association between metagenome and 16S rRNA genes:

Both the direct associations were sparse i.e. there were very few 16S reads present in each MGS Bin, while the indirect associations were not sparse. Hence, we intended to integrate the scores in a way that do not allow the indirect associations to dominate over the direct associations. For this, we converted these similarity scores to dissimilarity scores, combined them and revert them back to similarity scores, as done in STRING database (Szklarczyk et al., 2015). The only difference between the scoring scheme employed here and in STRING database (Szklarczyk et al., 2015) for

combining scores is that, we took a geometric mean of the dissimilarity scores while combining them instead of simply multiplying the different scores(as done in STRING database).

$$F = 1 - \sqrt[3]{(1-I)*(1-B)*(1-M)}$$

Where F= Combined score for the MGS Bins-16S rRNA genes relationship.

We observed that, the integrated correlation scores tend to dominate over the direct association scores in several instances. Hence, we decided not to take a geometric mean but to simply multiply the Pearson and Spearman correlation values to measure the consensus correlation value from both the methods:

$$V_{reg}(x,y) = value[I_{reg}(x,y)] = abs(P(x,y)) * abs(S(x,y))$$

$$I_{reg}(x,y) = V_{reg}(x,y) * Sg(x,y)$$

$$F_{reg} = 1 - \sqrt[3]{(1-I_{reg})*(1-B)*(1-M)}$$

Where

V_{reg}= Regularized integrated correlation value.

I_{reg} = Regularized integrated correlation score.

F_{reg} = Regularized combined score for the metagenome-16S rRNA genes relationship.

The closer the F_{reg} value to one, the higher the confidence of the metagenome-16S rRNA gene relationship. However, the 16S rRNA gene 'x' might have the highest confidence score to the MGS Bin 'y', but the MGS Bin 'y' need not have the highest confidence score to 16S rRNA gene 'x'. In order to tackle this scenario, we have to enrich these relationships by normalizing these scores by the highest confidence scores of the corresponding MGS Bin 'y' and 16S rRNA gene 'x'.

7.7.2.2.4 Enriching Metagenome bin to 16S rRNA gene relationship:

We estimated the probability of a MGS Bin 'y' to 16S rRNA gene 'x' relationship:

$$Pr\ (x,y) = \frac{F_{reg}(x,y)}{max_{0<x\leq m}\ F_{reg}(x,y)} * \frac{F_{reg}(x,y)}{max_{0<y\leq n}\ F_{reg}(x,y)}$$

The negative values of $Pr\ (x,y)$ could be turned zero and that could be considered an estimated probability of a MGS Bin 'y' belonging to a 16S rRNA gene 'x'. The normalized confidence score or the estimated probability thus obtained is the statistical likelihood of the confidence scores, adjusted for the background distribution of the confidence scores for all possible 16S-metagenome pair relationships.

7.7.2.3 Evaluation:

We evaluated the efficiency of the scoring scheme to link the MGS Bins to their corresponding 16S rRNA genes for two scenarios: 1) Metagenome bins possessing NCBI references. 2) Novel isolated strains with known MGS Bins and their corresponding 16S rRNA genes.

Scenario-1: As mentioned in the scenario above, if there were no traces of sequences constituting a 16S rRNA gene is available, then linking the 16S rRNA gene to the MGS Bin (representing the genome to which the 16S rRNA gene belong), is impossible. In such cases, the scoring scheme should be robust to link the MGS bins (whose 16S rRNA gene sequences were not present in the data) to a 16S rRNA gene with similar taxonomy annotation. For this, firstly we decided to identify the closest relative of the genome represented by an MGS Bin in NCBI database. Then, we evaluated whether the scoring scheme is able to link the MGS Bin to a 16S rRNA gene that is sequentially similar to that of the 16S rRNA gene present in the best-matched NCBI-Genome. To observe how the scoring scheme perform under such situations, we employed a strategy as depicted in **Figure 7-48**.

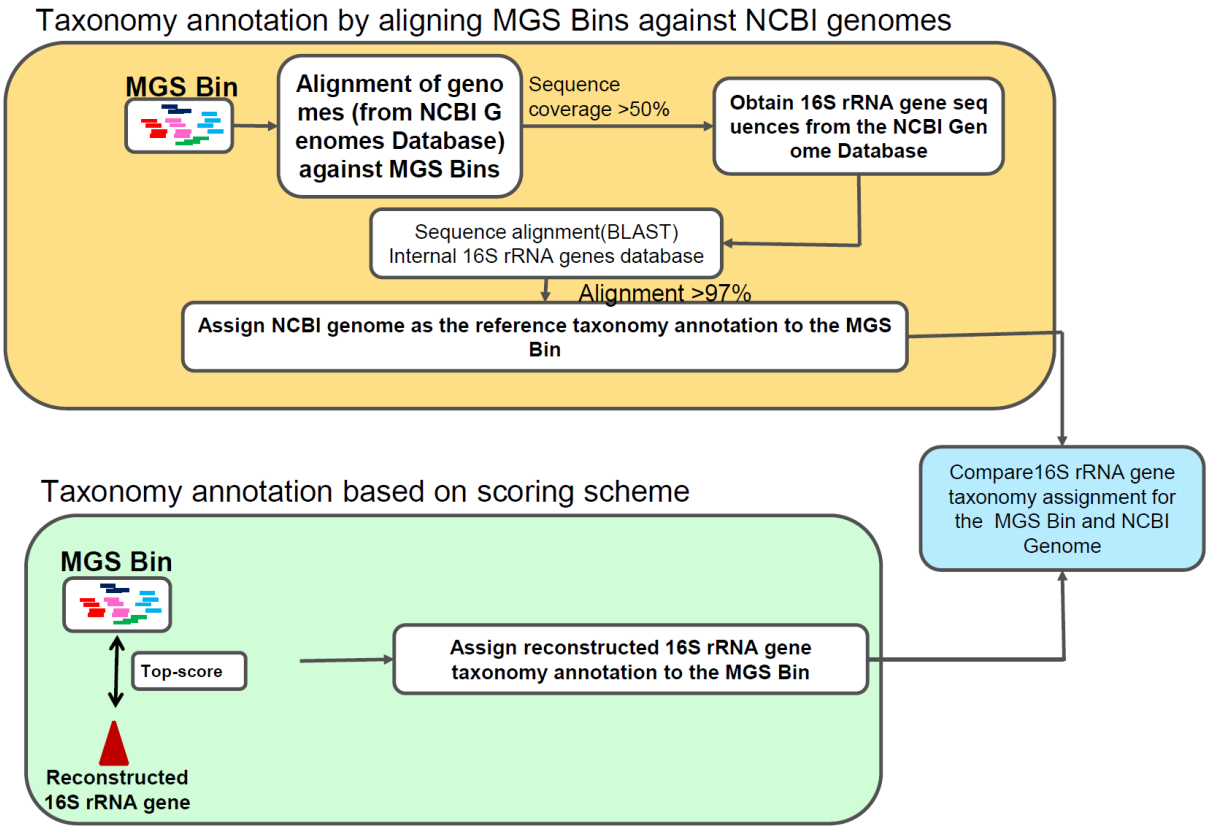


Figure 7-48 Workflow employed for evaluating the efficiency of the scoring scheme in 16S-MGS Bin linking, for Genomes from NCBI Database

The workflow employed for evaluating the taxonomy assigned to the MGS Bins via the scoring scheme, by comparing it against the taxonomy assigned to the same MGS Bin via matching them against genomes from NCBI Database. The yellow box represent the workflow employed to obtain taxonomy to the MGS Bin via matching them against genomes from NCBI Database. The green box represent the taxonomy assigned to the MGS Bin from the pool of reconstructed 16S rRNA genes via the scoring scheme developed in this study.

Briefly, we identified a genome from NCBI Database (an NCBI-Genome) that had high overall sequence similarity with an MGS Bin in our study, by aligning the contigs in each MGS Bin against NCBI Genomes. NCBI-reference: We considered those Genomes from NCBI Database that had a coverage >50% with an MGS Bin, to represent the closest relative of the MGS Bin in NCBI database. We picked the 16S rRNA gene sequence from this NCBI-Genome and mapped it against all the 16S rRNA genes reconstructed from all the MGS Bins in our study. We paired the reconstructed 16S rRNA gene, which had best match against the NCBI-16S rRNA gene sequence to the corresponding MGS bins and the taxonomy annotation of this particular reconstructed 16S rRNA gene as the reference taxonomy annotation for that particular MGS Bin. We observed whether the scoring scheme was also able to provide the highest confidence score for linking the reconstructed 16S rRNA gene and the MGS Bin. This is depicted by the **Figure 7-49**.

Taxonomy agreement: There were 47 MGS Bins had a match in NCBI Genome. 28 of the MGS bins were paired with a reconstructed 16S rRNA gene, both from the proposed scoring scheme and the NCBI-reference strategy and provided a consensus result. Amongst these 28 MGS Bins, we identified traces of 16S rRNA reads in 25 of them and 3 bins had no 16S reads in them. This shows that the scoring scheme was able to link the reconstructed 16S rRNA genes to their corresponding MGS Bins, even in scenarios where there were no 16S reads available in those MGS Bins and provide a consensus genus-level taxonomy annotation.

Taxonomy disagreement: However, both the methods disagreed in identifying the MGSBin-16S rRNA gene pairs for at least 19 MGS Bins. Amongst these, 12 MGS Bins were paired differently to 16S rRNA genes by both the methods, because we identified more number of 16S rRNA genes, in those MGS Bins not paired by the NCBI reference method. In spite of the difference in the MGS Bins-16S rRNA gene pairs by both the methods, 5 of the pairs linked by both the methods provided similar genus-level annotation, 4 of the pairs linked by both the methods provided similar family-level annotation, 1 pair obtained similar order-level annotation and only 1 pair similar kingdom-level annotation. We observed at least 7 MGS Bins were paired differently to 16S rRNA genes by both the methods where there were no 16S rRNA reads available in any of the sample. Even though both the methods disagreed, both the methods linked 2 of the MGS Bins to 16S rRNA genes of similar genus-level annotations. Amongst the rest, one of the pairs linked by both the methods provided similar family-level annotation, one of the pairs linked by both the methods provided similar order-level annotation, one of the pairs linked by both the methods provided similar family-level annotation, one of the pairs linked by both the methods provided similar class-level annotation, one of the pairs linked by both the methods provided similar phyla-level annotation and only 1 of the pairs linked by both the methods provided similar kingdom-level annotation.

$16S\ rRNA\ reads\ presence(columns)/Perfect\ Alignment(row$	Yes	No	Total
Yes	25	12	37
No	3	7	10
Total	28	19	47
Pearson's Chi-squared test with Yates' continuity correction : **p-value 0.0743**			

Table 4 Pearson's Exact test for identifying the significance of the number of 16S rRNA gene reads present and number of perfect alignment in results obtained from NCBI-reference and scoring scheme

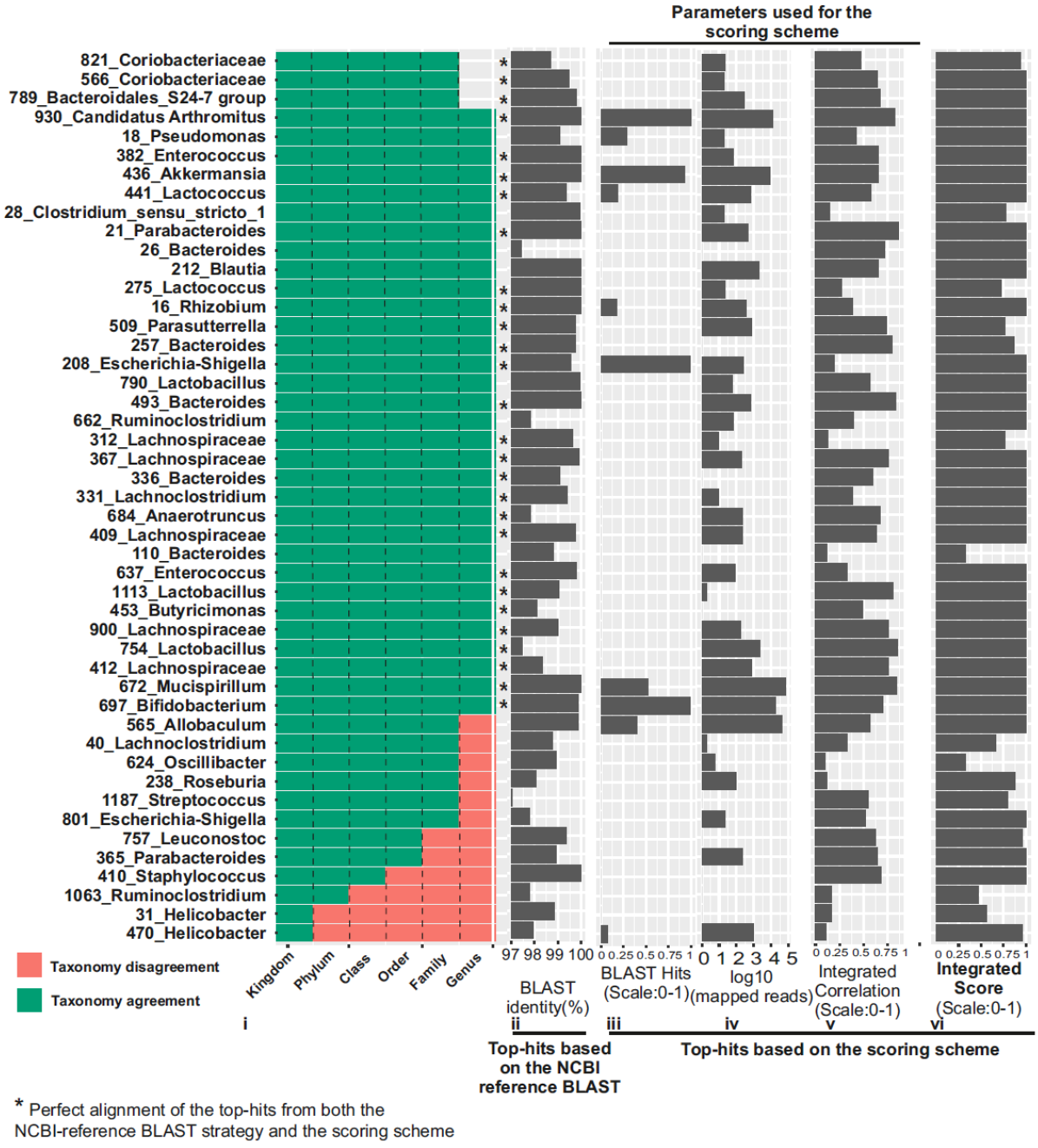


Figure 7-49 Comparison of taxonomic assignment between the scoring scheme and BLAST against genomes from NCBI Database.

(i) The bar plot represents the level of taxonomy alignment for bins whose 16S rRNA reads found matches in NCBI database and their corresponding taxonomy obtained by the scoring scheme described above. Red bar represents the level of taxonomy assigned based on NCBI database and green bar represent the level of taxonomy assigned based on the scoring scheme. (ii) The barplot on its right represents the identity (in percentage) of the 16S sequence inferred from the NCBI-Genome to the reconstructed 16S rRNA sequence of highest identity. (iii) The bar plot displays the number of 16S rRNA reads present in the MGS Bin with top-most integrated score, obtained by the linking strategy, as the barplot. (iv) We represented the number of 16S rRNA reads present in the MGS Bin with top-most integrated score, obtained by BLAST strategy, as the barplot. (iii) The bar plot displays the association between the 16S rRNA genes and the MGS Bin with the top-most integrated score, by the integrated correlation strategy, as the barplot. (iii) The bar plot displays the confidence score for linking the 16S rRNA genes and the MGS Bin with the top-most integrated score, as the barplot.

Scenario-2: To evaluate the efficiency of the scoring scheme developed to link 16S rRNA genes to the novel isolated strains of known taxonomy, we used an external set of novel bacterial strains whose 16S and genomes (Clavel et al., manuscript under preparation) were known. The results are depicted in .

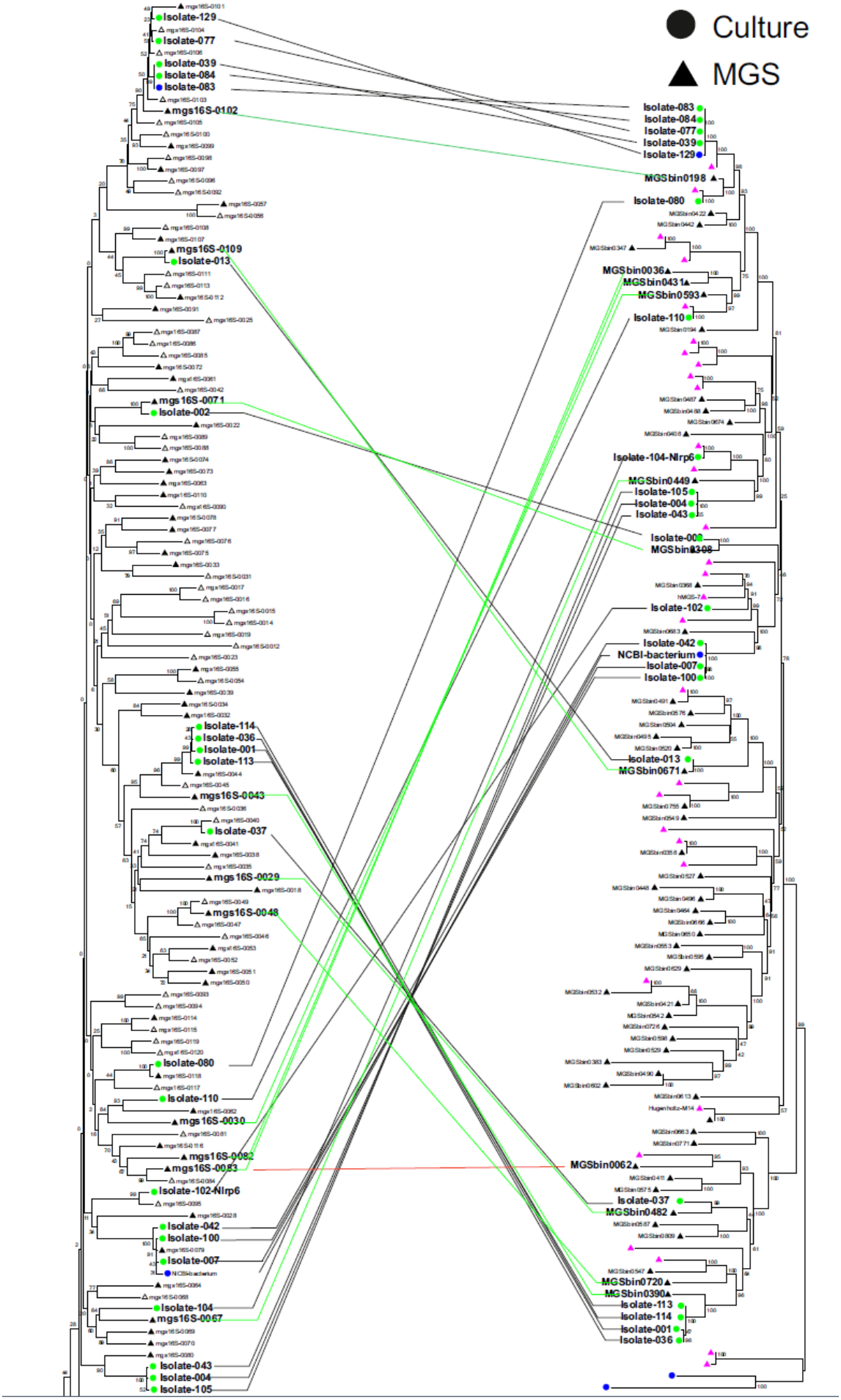


Figure 7-50 Evaluating the efficiency of scoring scheme for 16S-MGS Bin linking for novel isolated strains belonging to novel bacterial strains

The tree on the left represents the reconstructed 16S taxonomic tree for the novel bacterial strains and the tree on the right represents the reconstructed MGS Bins taxonomic tree for the novel bacterial strains. We knew the genomes to which each 16S rRNA genes belong, the black and green lines represent the scenarios where the scoring scheme prediction linked the 16S rRNA genes to the genome they actually belong, with the green lines specifying the mapping for the novel isolated novel bacterial strains. The red line represent represents the scenario where the scoring scheme prediction linked a 16S rRNA gene to the genome they do not belong.

The scoring scheme was able to provide a link between MGS Bin and an appropriate 16S rRNA gene for all the novel bacterial strains but one (indicated by the red line in **Figure 7-50).** Most importantly, the novel isolated novel bacterial strains (indicated by the green dots in the taxonomic trees) were assigned high scores to their corresponding 16S rRNA genes. This is feasible because most of the MGS Bins for Novel bacterial strains possessed at least a limited number of short sequences of the 16S rRNA genes in the metagenome data.

This indicates that the scoring scheme is robust to provide an appropriate taxonomic assignment for those MGS Bins possessing a small number of reads corresponding to the 16S rRNA gene.

This analysis indicate that the proposed novel scoring scheme to link the MGS Bins to their corresponding reconstructed 16S rRNA gene, with high confidence, is able to improve taxonomic resolution of the MGS Bins, though not a fool-proof model. The scoring scheme is mainly dependent on the identification of most of the 16S rRNA sequences corresponding to the MGS Bins under study. In case the taxonomic diversity of the 16S rRNA genes were only partially represented in the metagenomics study, it is highly difficult to reconstruct the 16S rRNA genes, thereby less chance of linking the MGS Bins to another 16S rRNA gene of similar high resolution taxonomy annotation. The evaluation of the performance of the scoring schemes have showed evidences for its ability to link MGS Bins with no 16S rRNA reads, to another reconstructed 16S rRNA gene with similar taxonomy annotation, thereby improving the resolution of the taxonomic assignment of the MGS Bins. However, further studies needs to be performed in this direction to identify the ways to regularize and improve the linking strategy, thereby improving the taxonomic resolution of the MGS Bins.

8 Discussion

The gut microbiota is an integral part of the barrier protecting healthy individuals from enteropathogenic infections. The members of the gut microbiota community help maintain intestinal homeostasis and contribute to many other physiological processes (Kamada et al., 201c). However if the microbiota composition is disturbed, pathogens like *S. Tm* might cause infections in susceptible individuals. *S. Tm* has a strong armory for sensing signals secreted by the host and the commensal bacteria and adapting to the gut environment by activating specific transcriptional machineries to survive and cause infection. Hence, an in-depth understanding of the specific role of the microbiota composition in an imbalanced gut environment to provide opportunity for pathogenesis is essential for the development of new and improved therapeutic agents for these infections. This project aims to understand the potential genetic signatures associated with the virulence and survival of *S*. Tm *in vivo* and the potential genetic interactions between the gut microbiota and *Salmonella* Typhimurium SL1344 (*S*. Tm). This study provides a unique opportunity to observe the *Salmonella* transcriptional machinery in response to varied murine gut microbial ecosystems at an early stage of their infection, by a transcriptome-metatranscriptome association study.

For this, murine models with different gut ecosystems (Germ-free, SPF-1 and SPF-2) were treated with streptomycin – to reduce protection and then infected with *S*. Tm. The RNA samples were collected 12 hours post infection from the cecal content. In addition, *S*. Tm was grown in aerobic and anaerobic conditions *in vitro.* We obtained the *S*. Tm expression profile by mapping the reads to its genome and the activity profile of the other members of the gut microbial community in SPF-1 and SPF-2 conditions by mapping them against the murine gut microbiota reference catalogue organized by Lesker et.al. (Manuscript under preparation). Based on these expression profiles, a) We characterized the transcriptional landscape of *S*. Tm in all the *in vitro* and *in vivo* conditions under study and identified the potential *S*. Tm gene signatures that could be associated with the host system and/ or the gut microbial community. b) We characterized the active members of the other members of the dysbiotic gut microbial ecosystem in SPF-1 and SPF-2 conditions and their potential functional systems that could be associated with *S*. Tm infection. c) We then performed transcriptome-metatranscriptome associative and interaction studies to identify the potential microbial signatures that could provide protection against *S*. Tm and *S*. Tm gene signatures employed to overcome resistance provided by the gut microbial ecosystem.

8.1 Advancing bioinformatic approaches for improved metatranscriptome analysis

8.1.1 Operon-based enrichment analysis for *Salmonella* transcriptome data

Based on the transcriptome data, we considered those *S*. Tm genetic elements that potentially showed difference in expression in response to the host-environment as host-associated signatures and those that showed difference in expression in response to microbial composition as microbiota-associated signatures. To analyze the *S*. Tm transcriptome data, we developed a heuristic approach that considered the operon structure for identifying *S*. Tm functional signatures associated with a specific condition. The conventional genome-wide transcriptome study of a microbe would involve identification of genes that were differentially expressed by comparing expression profiles of samples under two conditions. This is enabled by methods based on negative binomial distribution like DESeq2 (Love et al., 2014) and edgeR (Robinson et al., 2010). These methods are robust to detect differentially expressed genes with reduced false positive rates, yet the top genes obtained from these studies need not necessarily be functionally coordinated. To obtain the functionally coordinated set of genes that are highly/low expressed in one condition when compared to the other, Gene Set Enrichment Analysis methods like GSEA (Subramanian et al., 2005), npGSEA (Larson and Owen, 2015) are commonly employed. These methods provide us with an opportunity to identify the functionally coordinated genes showing concordant difference between in gene expression profiles two conditions. However, these analyses could be performed only on those genes with known functions. *S*. Tm has ~24% of the genes with unknown functions (genes with no functional annotations in Gene Ontology Biological Process, Gene Ontology Molecular Function and KEGG Orthology). Hence, gene-set enrichment analysis would ignore the expression profile of ~ 24% of the protein-coding genes. In addition, these Gene Set Enrichment Analysis methods do not perform well when the size of the gene sets were small. In order to tackle this issue, we employed a gene-set enrichment method that works as a wrapper to the results obtained by either DESeq2 or edgeR methods. Based on many previous studies, it has been showed that genes in an operons/TUs are mostly functionally coordinated (Salgado et al., 2000, Xiao et al., 2006, Wells et al., 2016) and most of the genes (~99%) in *S*. Tm belong to at least one operon. Hence, we organized the protein-coding genes to the Transcriptional Units (TUs)/operons they belong (3797 non-redundant and overlapping operons/TUs compiled from OperonDB (Pertea et al., 2009) and DOOR (Mao et al., 2009) database) and looked for the coordinated differential expression of these genes. We observed that the operon-based enrichment

analysis proposed in our study is able to classify the samples from different conditions better than the conventional DESeq2 method and random set of genes with same number of operons. This method also reduced false positive rates when compared to the conventional gene expression analysis for both in-house data and external data. A detailed description and evaluation of the method is provided in the Methods Section. Further improvements for the method should addressed for the following scenarios: a) It is difficult to obtain a complete operon structure in a novel isolated microbe. There are many tools available to predict the operon structure from the RNA-seq data like DOOR (Mao et al., 2009), however a careful investigation of such predicted operons needs to be performed before using them for any downstream analysis. In these situations, one could potentially use other alternative ways to organize the potential co-expressed sets-like KEGG Orthology or KEGG Modules could be done and our method of gene-set enrichment analysis could be employed, although there will be loss of information from ignoring genes with no functional annotations. b) We considered any operon with a fold change >2 and Kost's p-value <5% to be differentially expressed in samples between two conditions. Like any other method, setting up a threshold for the two parameters under consideration is key in identifying the functional systems characterizing the microbe under any given condition. The method could be improved in estimating the parameters (fold change and Kost's p-value) based on their ability to minimize intra-cluster distances and maximize inter-cluster distances.

Along with the enrichment analysis method, we have also proposed a logistic-regression based evaluation metric to identify whether the enriched operons/TUs were able to classify samples from different conditions better than any random set of operons/TUs. This heuristic metric use the coordinates obtained from Sammon distance (a distance metric that preserves the inter-sample distances while projecting a multi-dimensional data to two/three-dimensional data) to assess the quality of classification. Briefly, we organized S. Tm expression profile into operons/TUs they belong and normalized them using TPM method (Wagner et al., 2012). We obtained the Sammon distance between samples from any two conditions under study, based on the data normalized as described above. The coordinates obtained for the samples based on their Sammon distance values were modelled against the sample groups (mouse models) they belong. The quality of the model was assessed based on their Aikaike Information Criterion (AIC) score- the smaller the score, the better the quality of the model. We compared this AIC score against other models based on random set of operons/TUs. We observed that the operons/TUs selected based on the heuristic method developed in our study provided the best possible models. However, further

studies needs to be performed to assess the performance of this evaluation metric under different scenarios. We used the operon-based enrichment analysis method proposed in this study in identifying the gene signatures for all the comparative studies in our study.

8.1.2 Metagenome-16S linking strategy for improved taxonomic annotations

Also, the *S*. Tm-gut microbiota functional interaction identified in our study is feasible only due to the presence of the murine gut microbiota reference catalogue developed by Lesker et.al. (Manuscript under preparation) with improved taxonomic and functional annotations. To further improve the taxonomic resolution of the MGS Bins of this reference catalogue, we employed a strategy to reconstruct 16S rRNA genes and develop a scoring scheme to link the 16S rRNA gene with the corresponding MGS Bin, thereby assigning the taxonomic annotation of the 16S rRNA gene with improved taxonomic resolution to the linked MGS Bins with taxonomic annotation of higher level. The scoring scheme efficiently utilizes the sparse presence of the 16S rRNA gene sequences for better taxonomic assignment of the MGS Bins. However, the reconstructed 16S rRNA genes should be able to represent the taxonomic diversity of the microbial community and only then the scoring scheme could link the MGS Bins to 16S rRNA genes with similar taxonomic annotations. Though we observed the scoring scheme was robust in proving improved taxonomic annotations to novel isolated strains belonging to S24-7 groups and that of the NCBI genomes, further evaluation and regularization should be done to the scoring scheme based on simulation studies.

8.2 Insights into the interplay of *Salmonella* and the microbiota through metatranscriptome analysis

8.2.1 The adaptation of *Salmonella* to the gastrointestinal tract

In order to observe the key functional systems of *S*. Tm whose activation could be associated with the host system (i.e. to identify the potential host-associated gene signatures), we compared the *S*. Tm expression profile in the Germ-Free (GF) samples against those in the *in vitro* conditions. We observed that *S*. Tm employs a distinct transcriptional *in vivo* in comparison to its *in vitro* growth. We identified a) operons/TUs encoding uptake and utilization of known primary host-derived carbon sources of Salmonella in host system like propanediol (Sinha et al., 2012) and ethanolamine (Srikumar and Fuchs, 2011) b) SPI-1 system responsible for invasion of gut epithelial cells c). Operons encoding genes responsible for horizontal gene transfer to be highly expressed in Germ-Free samples in comparison to those samples grown *in vitro*

(aerobic and anaerobic conditions). In addition, we observed that operons/TUs encoding flagellar systems, membrane biosynthesis were downregulated in Germ-Free samples in comparison to those samples grown *in vitro* (aerobic and anaerobic conditions). Based on this we inferred that, in Germ-Free condition where there is no microbial competition, *S*. Tm readily feed on host-derived simple carbon sources and activates SPI-1 systems to cause infections. Also, *S*. Tm downregulated the membrane biosynthesis and flagellar system in order to evade the host immune system in identifying it as an antigenic target. We also observed that genes involved in TCA Cycle were also highly expressed in Germ-Free condition when compared to aerobic and anaerobic *in vitro* conditions. Many previous studies have indicated an association between the complete utilization of TCA Cycle and virulence in microbes of family *Enterobacteriaceae* - *Salmonella* Typhimurium *SR*-11 in BALB/c mice (Tchawa Yimga et al., 2006), and *Edwardsiella ictaluri* in catfish (Dahal et al., 2013). This shows us that there could be a strong association between the utilization of TCA Cycle and virulence under their favorable host conditions, at least in pathogens of family *Enterobacteriaceae.*

8.2.2 The influence of the microbiota on *Salmonella* gene expression

We compared the *Salmonella* expression profile in the Germ-Free samples against that in the colonized mouse models SPF-1 and SPF-2 conditions to identify the potential microbiota-associated gene signatures. Thiemann et al. (Thiemann et al., 2017) identified that the dysbiotic SPF-1 microbial ecosystem is more susceptible to *S*. Tm infection when compared to dysbiotic SPF-2 microbial ecosystem (in cecal tissue). This study gave us an opportunity to further characterize the molecular mechanisms of *S*. Tm during infection in these colonized mouse models. We observed that *S*. Tm maintains a core transcriptional machinery *in vivo,* with slight changes in them, attributed to few specific microenvironments created by a combinatorial effect of the gut microbiota composition and the host immune system. We observed that *S*. Tm upregulated operons/TUs responsible for uptake and utilization of multiple carbon sources in the colonized mouse models, in comparison to GF condition. However, citrate uptake and utilization systems were upregulated in GF condition in comparison to colonized mouse models. From this, we could infer that, in the presence of other microbiota and thereby competition to feed on primary simple carbon sources, *S*. Tm feeds on multiple, simple carbon sources all at once and derives energy from them. The operons/TUs responsible for performing TCA Cycle were highly expressed in all the *in vivo* samples in this study, evaluating the possible association between TCA

Cycle utilization and virulence, which was observed in *S*. Tm monocolonized with *Bacteroides thetaiotaomicron in vivo* (Spiga et al., 2017). Also, it was interesting to note that, feeding on multi-carbon sources could increase the sugar phosphate concentration before energy is derived from them. The analysis of the non-coding RNAs revealed that the small RNA SgrS, which is known to regulate sugar phosphate stress (Papenfort et al., 2012), was highly expressed in all the *in vivo* conditions. This indicate that, *S.* Tm transcribe the non-coding RNA SgrS responsible for regulating the sugar phosphate stress in all the *in vivo* conditions. In the absence of any competition, *S*. Tm feed on primary carbon sources like citrate to derive energy. From this, we could infer that the gut microbiota modulate the *S.* Tm transcriptional machinery, by modulating the microenvironments *S.* Tm encounter in the host gut environment and these microenvironments mainly include nutrient and metal-ion availability.

8.2.3 Metatranscriptome analysis

From the meta-transcriptome study, we aimed to identify the potential molecular mechanisms employed by the active members of the dysbiotic microbial community to outcompete other members for nutrient uptake, and the collective mechanism employed to protect or aid *S*. Tm infection. By doing so, we could identify the taxonomic groups and/or the functional groups associated with *Salmonella* virulence. The *Firmicutes* to *Bacteroides* ratio was comparatively low in SPF-2 condition in comparison to SPF-1 condition, which could be associated with the low susceptibility to *S*. Tm infection in SPF-2 in comparison to SPF-1 condition (in cecal tissue). In the SPF-1 condition, few species belonging to *Deferribacteria* phyla namely *Mucispirillum schaedleri* and *Firmicutes* phyla namely *Enterococcus faecelis*, *Ruminococcus bromii* and *Oscillibacter valericigenes* were the active members of the community. Amongst them, *Mucispirillum schaedleri* is a known pathobiont in the gut microbial community (Loy et al., 2017) and *Ruminococcus bromii* has genetic repertoire to degrade resistant starches (Ze et al., 2012). From this, one could infer that the higher susceptibility to S. Tm infection in SPF-1 condition could also be associated with the increased availability of simple carbon sources enabled by the activity of *Ruminococcus bromii* and/or the increased activity of the pathobiont *Mucispirillum schaedleri*. In SPF-2 condition, few species belonging to phyla *Firmicutes* namely *Enterococcus faecelis* and *Lachnospiracea bacterium A4* and species *Olsenella uli* belonging to phyla *Actinobacteria* were highly active. Amongst these active members of the SPF-2 microbial community, *Olsenella uli* is a lactic-acid related bacteria (LARB) and

Enterococcus faecalis is a lactic acid producing bacterium and few studies have indicated the protective role of lactic-acid bacteria against enteropathogens like *S.* Tm in different hosts namely chicken (Kim et al., 2015) and rats (Kim et al., 2013) highlighting the potential role of lactic-acid bacteria in providing protection against *S*. Tm in colonized mouse models.

8.2.4 *Salmonella*-microbiota association studies

In SPF-2 and SPF-1 *in vivo* conditions, *S*. T*m* respond to a combinatorial effect of both the host gut environment and the gut-microbiota activity. In order to identify the potential *S*. Tm-microbiota interactions, we performed an association study between the *S*. Tm expression profile and the metatranscriptome profile. Firstly, we identified the differentially regulated *S*. Tm transcriptional machineries in SPF-1 and SPF-2 conditions. Correspondingly, we identified the relatively active functional units in the SPF-1 and SPF-2 dysbiotic gut microbial communities and based on them, we identify the potential association between these *S.* Tm-microbiota signatures. Amongst the functional systems that were highly expressed in SPF-1 microbial community in comparison to SPF-2 microbial community, we observed TCA Cycle and Bacterial Secretion System were highly expressed, similar to which *S.* Tm also upregulated genes responsible for TCA Cycle in the SPF-1 condition. This evaluate the possible association between the utilization of TCA Cycle and virulence that we observed in *S*. Tm *in vivo* (as mentioned previously). The gut microbiota in SPF-1 condition showed higher expression of Type-II Secretion System that are capable of translocating enzymes like proteases and chitinases (Cianciotto and White, 2017) that could aid *Salmonella* in feeding on diet-derived disaccharide like cellobiose and host-derived monosaccharide like sialic acid. Correspondingly, cellobiose and sialic acid uptake systems were upregulated in *S.* Tm in SPF-1 condition. In SPF-2 condition, both *S. Tm* and gut microbiota ecosystem showed far less number of genes involved in central carbon metabolism and stress responses upregulated in comparison to SPF-1 condition. This could indicate either reduced nutrient availability or high nutrient competition for *S*. Tm in SPF-2 condition. This section allowed us in identifying the potential association between the ability of *S*. Tm in feeding primary simple carbon sources and its virulence. We could also infer that the gut microbiota could influence the severity of *S*. Tm virulence by providing competition to feed on the primary carbon sources.

8.2.5 *Salmonella*-microbiota functional interaction studies

Even though association studies allowed us to identify the association between S. Tm and gut microbiota functional signatures, we were unable to identify the potential active member expressing these functional signatures. For this, we employed a novel framework for *S*. Tm-microbiota functional interactions. By implementing this framework we aimed in identifying the potential *S*. Tm-microbial functional interactions in SPF-1 and SPF-2 conditions. For this, we identified those functional systems that were active in *S.* Tm*,* but not in the gut microbiota and vice versa.

8.2.5.1 Microbial functional signatures associated to *S*. Tm virulence

First, we identified the functional signatures that were expressed higher than *S*. Tm, by the gut microbial community that could be associated with *S.* Tm virulence. In the SPF-1 condition, we observed functional system responsible for succinyl-CoA production were highly expressed by the gut microbial community. A previous study by Spiga et al. (Spiga et al., 2017) has indicated that *S*. Tm feeds on microbiota-derived succinate and causes infection, which evaluated the observation in our study. From this, we could also infer that the microbiota-derived succinyl-CoA could potentially aid *S*. Tm infection in the SPF-1 condition. It was striking to observe that most of the active members of the SPF-1 gut microbial community had the ability to encode such functional systems. In the SPF-2 condition, a recently discovered energy coupling factor transport system associated with uptake of multiple micronutrients like biotin, riboflavin and thiamine (Karpowich and Wang, 2013) were highly expressed by the gut microbial community. Even in this scenario, we observed that most of the active members of the SPF-2 gut microbial community had the gene repertoire to encode such functional systems. This could suggest the significance of the redundancy of such functional signatures in the genetic repertoire of many of the active members of the community, which was observed in (Kang et al., 2015), unlike species redundancy or the influence of keystone species.

8.2.5.2 *S.* Tm functional associated to its *in vivo* fitness

Correspondingly, we identified the *S.* Tm functional signatures that were expressed higher than the gut microbial community that could be associated with *S.* Tm virulence. We observed multi-drug resistance, nitrate respiration, T1SS, RTX toxin system and C3-C4 carbon source uptake systems were highly expressed in *S*. Tm when compared to other member of the SPF-1 gut microbial community. This indicates that *S*. Tm has

a robust transcriptional machinery to adapt and survive in host gut system. The active members of SPF-2 microbial community had the gene repertoire to perform most of these functions, especially those belong to *Lachnospiraceae*. Based on this, we could hypothesis that the active members of the SPF-2 community might possess functional potentials that could provide competition in micronutrients acquisition to *S*. Tm thereby reducing its infection. This also showed us that the novel *S*. Tm-microbiota interaction framework introduced in our study allowed us to understand and identify the microbiota-modulated *S*. Tm transcriptional machinery in better resolution when compared to that of the conventional *S*. Tm-microbiota association studies. However, this hypothesis should be evaluated experimentally in order to identify the potential microbe and the potential functional signature that is directly responsible for increase in protection against *S*. Tm in host gut environment. Correspondingly, in the SPF-2 community, we observed that *S*. Tm highly expressed its T3SS injectisome, which was not in the genetic repertoire of any member of the SPF-1 or SPF-2 gut microbial community. There have been many recent attempts to inhibit the T3SS (Gu et al., 2015, (Li et al., 2013), Duncan et al., 2012; McShan and De Guzman, 2015) for developing anti-bacterial agents against the gram-negative pathogens. These observations suggest that *S*. Tm activity in host gut environment were greatly influenced by the microbiota composition and their functional potentials in providing protection against *S*. Tm. Also, the current study provided a 'snap-shot' of the transcriptional landscapes of *S*. Tm and gut microbiota. The observations made in the study should be evaluated by performing a temporal study of the transcriptional landscape of *S*. Tm and gut microbiota.

9 Reference

Altuvia, S., Weinstein-Fischer, D., Zhang, A., Postow, L., and Storz, G. (1997). A small, stable RNA induced by oxidative stress: role as a pleiotropic regulator and antimutator. Cell *90*, 43–53.

Baba, Y., Iwatsuki, M., Yoshida, N., Watanabe, M., and Baba, H. (2017). Review of the gut microbiome and esophageal cancer: Pathogenesis and potential clinical implications. Ann. Gastroenterol. Surg. *1*, 99–104.

Baez, S., and Gordon, H.A. (1971). Tone and Reactivity of Vascular Smooth Muscle in Germfree Rat Mesentery. J. Exp. Med. *134*, 846–856.

Barchiesi, J., Espariz, M., Checa, S.K., and Soncini, F.C. (2009). Downregulation of RpoN-controlled genes protects Salmonella cells from killing by the cationic antimicrobial peptide polymyxin B. FEMS Microbiol. Lett. *291*, 73–79.

Barthel, M., Hapfelmeier, S., Quintanilla-Martínez, L., Kremer, M., Rohde, M., Hogardt, M., Pfeffer, K., Rüssmann, H., and Hardt, W.-D. (2003). Pretreatment of Mice with Streptomycin Provides a Salmonella enterica Serovar Typhimurium Colitis Model That Allows Analysis of Both Pathogen and Host. Infect. Immun. *71*, 2839–2858.

Bäumler, A.J. (1997). The record of horizontal gene transfer in Salmonella. Trends Microbiol. *5*, 318–322.

Belkaid, Y., and Hand, T. (2014). Role of the Microbiota in Immunity and inflammation. Cell *157*, 121–141.

Bengmark, S. (2013). Gut microbiota, immune development and function. Pharmacol. Res. *69*, 87–113.

Bertani, G. (2004). Lysogeny at Mid-Twentieth Century: P1, P2, and Other Experimental Systems. J. Bacteriol. *186*, 595–600.

Biegel, E., Schmidt, S., González, J.M., and Müller, V. (2011). Biochemistry, evolution and physiological function of the Rnf complex, a novel ion-motive electron transport complex in prokaryotes. Cell. Mol. Life Sci. *68*, 613–634.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Bowden, S.D., Ramachandran, V.K., Knudsen, G.M., Hinton, J.C.D., and Thompson, A. (2010). An Incomplete TCA Cycle Increases Survival of Salmonella Typhimurium during Infection of Resting and Activated Murine Macrophages. PLOS ONE *5*, e13871.

Brugiroux, S., Beutler, M., Pfann, C., Garzetti, D., Ruscheweyh, H.-J., Ring, D., Diehl, M., Herp, S., Lötscher, Y., Hussain, S., et al. (2016). Genome-guided design of a defined mouse microbiota that confers colonization resistance against *Salmonella enterica* serovar Typhimurium. Nat. Microbiol. *2*, nmicrobiol2016215.

Buffie, C.G., and Pamer, E.G. (2013). Microbiota-mediated colonization resistance against intestinal pathogens. Nat. Rev. Immunol. *13*, 790–801.

Bull, M.J., and Plummer, N.T. (2014). Part 1: The Human Gut Microbiome in Health and Disease. Integr. Med. Clin. J. *13*, 17–22.

Campos, E., de la Riva, L., Garces, F., Giménez, R., Aguilar, J., Baldoma, L., and Badia, J. (2008). The yiaKLX1X2PQRS and ulaABCDEFG Gene Systems Are Required for the Aerobic Utilization of l-Ascorbate in Klebsiella pneumoniae Strain 13882 with l-Ascorbate-6-Phosphate as the Inducer. J. Bacteriol. *190*, 6615–6624.

Chassard, C., and Lacroix, C. (2013). Carbohydrates and the human gut microbiota. Curr. Opin. Clin. Nutr. Metab. Care *16*, 453–460.

Chaudhuri, R.R., Morgan, E., Peters, S.E., Pleasance, S.J., Hudson, D.L., Davies, H.M., Wang, J., Diemen, P.M. van, Buckley, A.M., Bowen, A.J., et al. (2013). Comprehensive Assignment of Roles for Salmonella Typhimurium Genes in Intestinal Colonization of Food-Producing Animals. PLOS Genet. *9*, e1003456.

Cheng, S., and Bobik, T.A. (2010). Characterization of the PduS Cobalamin Reductase of Salmonella enterica and Its Role in the Pdu Microcompartment. J. Bacteriol. *192*, 5071–5080.

Choi, J., Shin, D., Yoon, H., Kim, J., Lee, C.-R., Kim, M., Seok, Y.-J., and Ryu, S. (2010). Salmonella pathogenicity island 2 expression negatively controlled by EIIANtr–SsrB interaction is required for Salmonella virulence. Proc. Natl. Acad. Sci. *107*, 20506–20511.

Christie, P.J., Whitaker, N., and González-Rivera, C. (2014). Mechanism and structure of the bacterial type IV secretion systems. Biochim. Biophys. Acta BBA - Mol. Cell Res. *1843*, 1578–1591.

Chung, H., Pamp, S.J., Hill, J.A., Surana, N.K., Edelman, S.M., Troy, E.B., Reading, N.C., Villablanca, E.J., Wang, S., Mora, J.R., et al. (2012). Gut immune maturation depends on colonization with a host-specific microbiota. Cell *149*, 1578–1593.

Cianciotto, N.P., and White, R.C. (2017). Expanding Role of Type II Secretion in Bacterial Pathogenesis and Beyond. Infect. Immun. *85*.

Clarke, S.F., Murphy, E.F., Nilaweera, K., Ross, P.R., Shanahan, F., O'Toole, P.W., and Cotter, P.D. (2012). The gut microbiota and its relationship to diet and obesity. Gut Microbes *3*, 186–202.

Collins, S.M., and Bercik, P. (2009). The Relationship Between Intestinal Microbiota and the Central Nervous System in Normal Gastrointestinal Function and Disease. Gastroenterology *136*, 2003–2014.

Conly, J.M., and Stein, K. (1992). The production of menaquinones (vitamin K2) by intestinal bacteria and their role in maintaining coagulation homeostasis. Prog. Food Nutr. Sci. *16*, 307–343.

Crawford, P.A., Crowley, J.R., Sambandam, N., Muegge, B.D., Costello, E.K., Hamady, M., Knight, R., and Gordon, J.I. (2009). Regulation of myocardial ketone body metabolism by the gut microbiota during nutrient deprivation. Proc. Natl. Acad. Sci. *106*, 11276–11281.

Dahal, N., Abdelhamed, H., Lu, J., Karsi, A., and Lawrence, M.L. (2013). Tricarboxylic Acid Cycle and One-Carbon Metabolism Pathways Are Important in Edwardsiella ictaluri Virulence. PLOS ONE *8*, e65973.

De Lay, N., and Gottesman, S. (2012). A complex network of small non-coding RNAs regulate motility in Escherichia coli. Mol. Microbiol. *86*, 524–538.

Dieye, Y., Ameiss, K., Mellata, M., and Curtiss, R. (2009). The Salmonella Pathogenicity Island (SPI) 1 contributes more than SPI2 to the colonization of the chicken by Salmonella enterica serovar Typhimurium. BMC Microbiol. *9*, 3.

Duncan, M.C., Linington, R.G., and Auerbuch, V. (2012). Chemical Inhibitors of the Type Three Secretion System: Disarming Bacterial Pathogens. Antimicrob. Agents Chemother. *56*, 5433–5441.

Eteleeb, A. (2015). An island-based approach for RNA-SEQ differential expression analysis. Electron. Theses Diss.

Eteleeb, A.M., Flight, R.M., Harrison, B.J., Petruska, J.C., and Rouchka, E.C. (2013). An Island-Based Approach for Differential Expression Analysis. 2013 ACM Conf. Bioinforma. Comput. Biol. Biomed. Inform. ACM - BCB 2013 Wash. DC USA Sept. 22 - 25 2013 Cathy Wu Sridhar Hannenhalli Gen. Co-Chairs ACM Conf. Bioinf *2013*, 419–429.

Fàbrega, A., and Vila, J. (2013). Salmonella enterica Serovar Typhimurium Skills To Succeed in the Host: Virulence and Regulation. Clin. Microbiol. Rev. *26*, 308–341.

Faith, J.J., McNulty, N.P., Rey, F.E., and Gordon, J.I. (2011). Predicting a human gut microbiota's response to diet in gnotobiotic mice. Science *333*, 101–104.

Flint, H.J., Bayer, E.A., Rincon, M.T., Lamed, R., and White, B.A. (2008). Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis. Nat. Rev. Microbiol. *6*, 121–131.

Fookes, M., Schroeder, G.N., Langridge, G.C., Blondel, C.J., Mammina, C., Connor, T.R., Seth-Smith, H., Vernikos, G.S., Robinson, K.S., Sanders, M., et al. (2011). Salmonella bongori Provides Insights into the Evolution of the Salmonellae. PLoS Pathog. *7*.

Galán, J.E. (2001). Salmonella Interactions with Host Cells: Type III Secretion at Work. Annu. Rev. Cell Dev. Biol. *17*, 53–86.

Garriga, X., Eliasson, R., Torrents, E., Jordan, A., Barbé, J., Gibert, I., and Reichard, P. (1996). nrdD and nrdG genes are essential for strict anaerobic growth of Escherichia coli. Biochem. Biophys. Res. Commun. *229*, 189–192.

Gerstle, K., Klätschke, K., Hahn, U., and Piganeau, N. (2012). The small RNA RybA regulates key-genes in the biosynthesis of aromatic amino acids under peroxide stress in E. coli. RNA Biol. *9*, 458–468.

Gilbreath, J.J., Colvocoresses Dodds, J., Rick, P.D., Soloski, M.J., Merrell, D.S., and Metcalf, E.S. (2012). Enterobacterial Common Antigen Mutants of Salmonella enterica Serovar Typhimurium Establish a Persistent Infection and Provide Protection against Subsequent Lethal Challenge. Infect. Immun. *80*, 441–450.

Goldman, B.S., Gabbert, K.K., and Kranz, R.G. (1996). The temperature-sensitive growth and survival phenotypes of Escherichia coli cydDC and cydAB strains are due to deficiencies in cytochrome bd and are corrected by exogenous catalase and reducing agents. J. Bacteriol. *178*, 6348–6351.

Gu, L., Zhou, S., Zhu, L., Liang, C., and Chen, X. (2015). Small-Molecule Inhibitors of the Type III Secretion System. Mol. Basel Switz. *20*, 17659–17674.

Hawrelak, J.A., and Myers, S.P. (2004). The causes of intestinal dysbiosis: a review. Altern. Med. Rev. J. Clin. Ther. *9*, 180–197.

Hensel, M. (2000). Salmonella Pathogenicity Island 2. Mol. Microbiol. *36*, 1015–1023.

Hensel, M., Hinsley, A.P., Nikolaus, T., Sawers, G., and Berks, B.C. (1999). The genetic basis of tetrathionate respiration in Salmonella typhimurium. Mol. Microbiol. *32*, 275–287.

Hoffmann, C., Hill, D.A., Minkah, N., Kirn, T., Troy, A., Artis, D., and Bushman, F. (2009). Community-Wide Response of the Gut Microbiota to Enteropathogenic Citrobacter rodentium Infection Revealed by Deep Sequencing. Infect. Immun. *77*, 4668–4678.

Ibañez, E., Campos, E., Baldoma, L., Aguilar, J., and Badia, J. (2000). Regulation of Expression of the yiaKLMNOPQRS Operon for Carbohydrate Utilization in Escherichia coli: Involvement of the Main Transcriptional Factors. J. Bacteriol. *182*, 4617–4624.

Iuchi, S., Chepuri, V., Fu, H.A., Gennis, R.B., and Lin, E.C. (1990). Requirement for terminal cytochromes in generation of the aerobic signal for the arc regulatory system in Escherichia coli: study utilizing deletions and lac fusions of cyo and cyd. J. Bacteriol. *172*, 6020–6025.

Jandhyala, S.M., Talukdar, R., Subramanyam, C., Vuyyuru, H., Sasikala, M., and Reddy, D.N. (2015). Role of the normal gut microbiota. World J. Gastroenterol. WJG *21*, 8787–8803.

Kamada, N., Chen, G.Y., Inohara, N., and Núñez, G. (2013). Control of pathogens and pathobionts by the gut microbiota. Nat. Immunol. *14*, 685–690.

Kang, S., Ma, W., Li, F.Y., Zhang, Q., Niu, J., Ding, Y., Han, F., and Sun, X. (2015). Functional Redundancy Instead of Species Redundancy Determines Community Stability in a Typical Steppe of Inner Mongolia. PLoS ONE *10*.

Karpowich, N.K., and Wang, D.-N. (2013). Assembly and mechanism of a group II ECF transporter. Proc. Natl. Acad. Sci. U. S. A. *110*, 2534–2539.

Kim, J.Y., Young, J.A., Gunther, N.W., and Lee, J.-L. (2015). Inhibition of Salmonella by Bacteriocin-Producing Lactic Acid Bacteria Derived from U.S. Kimchi and Broiler Chicken. J. Food Saf. *35*, 1–12.

Kim, M.-S., Yoon, Y.-S., Seo, J.-G., Lee, H.-G., Chung, M.-J., and Yum, D.-Y. (2013). A Study on the Prevention of Salmonella Infection by Using the Aggregation Characteristics of Lactic Acid Bacteria. Toxicol. Res. *29*, 129–135.

Koliada, A., Syzenko, G., Moseiko, V., Budovska, L., Puchkov, K., Perederiy, V., Gavalko, Y., Dorofeyev, A., Romanenko, M., Tkach, S., et al. (2017). Association between body mass index and Firmicutes/Bacteroidetes ratio in an adult Ukrainian population. BMC Microbiol. *17*.

Kopylova, E., Noé, L., and Touzet, H. (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinforma. Oxf. Engl. *28*, 3211–3217.

Kost, J.T., and McDermott, M.P. (2002). Combining dependent P-values. Stat. Probab. Lett. *60*, 183–190.

Kostic, A.D., Howitt, M.R., and Garrett, W.S. (2013). Exploring host–microbiota interactions in animal models and humans. Genes Dev. *27*, 701–718.

Kröger, C., Dillon, S.C., Cameron, A.D.S., Papenfort, K., Sivasankaran, S.K., Hokamp, K., Chao, Y., Sittka, A., Hébrard, M., Händler, K., et al. (2012). The transcriptional landscape and small RNAs of Salmonella enterica serovar Typhimurium. Proc. Natl. Acad. Sci. U. S. A. *109*, E1277-1286.

Kröger, C., Colgan, A., Srikumar, S., Händler, K., Sivasankaran, S.K., Hammarlöf, D.L., Canals, R., Grissom, J.E., Conway, T., Hokamp, K., et al. (2013). An Infection-Relevant Transcriptomic Compendium for Salmonella enterica Serovar Typhimurium. Cell Host Microbe *14*, 683–695.

Kumar, R., and Shimizu, K. (2011). Transcriptional regulation of main metabolic pathways of cyoA, cydB, fnr, and fur gene knockout Escherichia coli in C-limited and N-limited aerobic continuous cultures. Microb. Cell Factories *10*, 3.

Lamichhane-Khadka, R., Frye, J.G., Porwollik, S., McClelland, M., and Maier, R.J. (2011). Hydrogen-Stimulated Carbon Acquisition and Conservation in Salmonella enterica Serovar Typhimurium ▽ . J. Bacteriol. *193*, 5824–5832.

Larson, J.L., and Owen, A.B. (2015). Moment based gene set tests. BMC Bioinformatics *16*, 132.

Ley, R.E., Bäckhed, F., Turnbaugh, P., Lozupone, C.A., Knight, R.D., and Gordon, J.I. (2005). Obesity alters gut microbial ecology. Proc. Natl. Acad. Sci. U. S. A. *102*, 11070–11075.

Ley, R.E., Turnbaugh, P.J., Klein, S., and Gordon, J.I. (2006). Microbial ecology: Human gut microbes associated with obesity. Nature *444*, 1022.

Li, J., Lv, C., Sun, W., Li, Z., Han, X., Li, Y., and Shen, Y. (2013). Cytosporone B, an Inhibitor of the Type III Secretion System of Salmonella enterica Serovar Typhimurium. Antimicrob. Agents Chemother. *57*, 2191–2198.

Liu, M.Y., and Romeo, T. (1997). The global regulator CsrA of Escherichia coli is a specific mRNA-binding protein. J. Bacteriol. *179*, 4639–4642.

Liu, M.Y., Gui, G., Wei, B., Preston, J.F., Oakford, L., Yüksel, Ü., Giedroc, D.P., and Romeo, T. (1997). The RNA Molecule CsrB Binds to the Global Regulatory Protein CsrA and Antagonizes Its Activity in Escherichia coli. J. Biol. Chem. *272*, 17502–17510.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Loy, A., Pfann, C., Steinberger, M., Hanson, B., Herp, S., Brugiroux, S., Neto, J.C.G., Boekschoten, M.V., Schwab, C., Urich, T., et al. (2017). Lifestyle and Horizontal Gene Transfer-Mediated Evolution of Mucispirillum schaedleri, a Core Member of the Murine Gut Microbiota. MSystems *2*, e00171-16.

Maier, R.J. (2005). Use of molecular hydrogen as an energy substrate by human pathogenic bacteria. Biochem. Soc. Trans. *33*, 83–85.

Mao, F., Dam, P., Chou, J., Olman, V., and Xu, Y. (2009). DOOR: a database for prokaryotic operons. Nucleic Acids Res. *37*, D459-463.

Mardinoglu, A., Shoaie, S., Bergentall, M., Ghaffari, P., Zhang, C., Larsson, E., Bäckhed, F., and Nielsen, J. (2015). The gut microbiota modulates host amino acid and glutathione metabolism in mice. Mol. Syst. Biol. *11*, 834.

Marshall, J.M., and Gunn, J.S. (2015). The O-Antigen Capsule of Salmonella enterica Serovar Typhimurium Facilitates Serum Resistance and Surface Expression of FliC. Infect. Immun. *83*, 3946–3959.

Martens, J.-H., Barg, H., Warren, M., and Jahn, D. (2002). Microbial production of vitamin B12. Appl. Microbiol. Biotechnol. *58*, 275–285.

McClure, R., Balasubramanian, D., Sun, Y., Bobrovskyy, M., Sumby, P., Genco, C.A., Vanderpool, C.K., and Tjaden, B. (2013). Computational analysis of bacterial RNA-Seq data. Nucleic Acids Res. *41*, e140.

McShan, A.C., and De Guzman, R.N. (2015). The Bacterial Type III Secretion System as a Target for Developing New Antibiotics. Chem. Biol. Drug Des. *85*, 30–42.

Mercado-Lubo, R., Gauger, E.J., Leatham, M.P., Conway, T., and Cohen, P.S. (2008). A Salmonella enterica serovar typhimurium succinate dehydrogenase/fumarate reductase double mutant is avirulent and immunogenic in BALB/c mice. Infect. Immun. *76*, 1128–1134.

Mikheenko, A., Saveliev, V., and Gurevich, A. (2016). MetaQUAST: evaluation of metagenome assemblies. Bioinformatics *32*, 1088–1090.

Mirold, S., Ehrbar, K., Weissmüller, A., Prager, R., Tschäpe, H., Rüssmann, H., and Hardt, W.D. (2001). Salmonella host cell invasion emerged by acquisition of a mosaic of separate genetic elements, including Salmonella pathogenicity island 1 (SPI1), SPI5, and sopE2. J. Bacteriol. *183*, 2348–2358.

Mølbak, K., Baggesen, D.L., Aarestrup, F.M., Ebbesen, J.M., Engberg, J., Frydendahl, K., Gerner-Smidt, P., Petersen, A.M., and Wegener, H.C. (1999). An outbreak of multidrug-resistant, quinolone-resistant Salmonella enterica serotype typhimurium DT104. N. Engl. J. Med. *341*, 1420–1425.

Monterrubio, R., Baldoma, L., Obradors, N., Aguilar, J., and Badia, J. (2000). A Common Regulator for the Operons Encoding the Enzymes Involved in d-Galactarate, d-Glucarate, and d-Glycerate Utilization in Escherichia coli. J. Bacteriol. *182*, 2672–2674.

Morris, C., Yip, C.M.C., Tsui, I.S.M., Wong, D.K.-H., and Hackett, J. (2003). The Shufflon of Salmonella enterica Serovar Typhi Regulates Type IVB Pilus-Mediated Bacterial Self-Association. Infect. Immun. *71*, 1141–1146.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods *5*, nmeth.1226.

Nguyen, T.L.A., Vieira-Silva, S., Liston, A., and Raes, J. (2015). How informative is the mouse for human gut microbiota research? Dis. Model. Mech. *8*, 1–16.

Nishiyama, S., Murakami, Y., Nagata, H., Shizukuishi, S., Kawagishi, I., and Yoshimura, F. (2007). Involvement of minor components associated with the FimA fimbriae of Porphyromonas gingivalis in adhesive functions. Microbiol. Read. Engl. *153*, 1916–1925.

Ortega, Á.D., Gonzalo-Asensio, J., and Portillo, F.G. (2012). Dynamics of Salmonella small RNA expression in non-growing bacteria located inside eukaryotic cells. RNA Biol. *9*, 469–488.

Oshota, O., Conway, M., Fookes, M., Schreiber, F., Chaudhuri, R.R., Yu, L., Morgan, F.J.E., Clare, S., Choudhary, J., Thomson, N.R., et al. (2017). Transcriptome and proteome analysis of Salmonella enterica serovar Typhimurium systemic infection of wild type and immune-deficient mice. PLoS ONE *12*.

Pachter, L. (2011). Models for transcript quantification from RNA-Seq. ArXiv11043889 Q-Bio Stat.

Paiva, J.B., Penha Filho, R. a. C., Pereira, E.A., Lemos, M.V.F., Barrow, P.A., Lovell, M.A., and Berchieri, A. (2009). The contribution of genes required for anaerobic respiration to the virulence of Salmonella enterica serovar Gallinarum for chickens. Braz. J. Microbiol. Publ. Braz. Soc. Microbiol. *40*, 994–1001.

de Paiva, J.B., Penha Filho, R.A.C., Arguello, Y.M.S., Berchieri Junior, Â., Lemos, M.V.F., and Barrow, P.A. (2009). A defective mutant of Salmonella enterica Serovar Gallinarum in cobalamin biosynthesis is avirulent in chickens. Braz. J. Microbiol. *40*, 495–504.

Palaniyandi, S., Mitra, A., Herren, C.D., Lockatell, C.V., Johnson, D.E., Zhu, X., and Mukhopadhyay, S. (2012). BarA-UvrY Two-Component System Regulates Virulence of Uropathogenic E. coli CFT073. PLoS ONE *7*.

Papenfort, K., Podkaminski, D., Hinton, J.C.D., and Vogel, J. (2012). The ancestral SgrS RNA discriminates horizontally acquired Salmonella mRNAs through a single G-U wobble pair. Proc. Natl. Acad. Sci. U. S. A. *109*, E757–E764.

Pertea, M., Ayanbule, K., Smedinghoff, M., and Salzberg, S.L. (2009). OperonDB: a comprehensive database of predicted operons in microbial genomes. Nucleic Acids Res. *37*, D479–D482.

Ramachandran, V.K., Shearer, N., Jacob, J.J., Sharma, C.M., and Thompson, A. (2012). The architecture and ppGpp-dependent expression of the primary transcriptome of Salmonella Typhimurium during invasion gene expression. BMC Genomics *13*, 25.

Rangan, K.J., Pedicord, V.A., Wang, Y.-C., Kim, B., Lu, Y., Shaham, S., Mucida, D., and Hang, H.C. (2016). A secreted bacterial peptidoglycan hydrolase enhances tolerance to enteric pathogens. Science *353*, 1434–1437.

Rey, F.E., Faith, J.J., Bain, J., Muehlbauer, M.J., Stevens, R.D., Newgard, C.B., and Gordon, J.I. (2010). Dissecting the in vivo metabolic potential of two human gut acetogens. J. Biol. Chem. *285*, 22082–22090.

Rhee, S.H., Pothoulakis, C., and Mayer, E.A. (2009). Principles and clinical implications of the brain–gut–enteric microbiota axis. Nat. Rev. Gastroenterol. Hepatol. *6*.

Ridaura, V.K., Faith, J.J., Rey, F.E., Cheng, J., Duncan, A.E., Kau, A.L., Griffin, N.W., Lombard, V., Henrissat, B., Bain, J.R., et al. (2013). Cultured gut microbiota from twins discordant for obesity modulate adiposity and metabolic phenotypes in mice. Science *341*.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

Rodríguez, J.M., Murphy, K., Stanton, C., Ross, R.P., Kober, O.I., Juge, N., Avershina, E., Rudi, K., Narbad, A., Jenmalm, M.C., et al. (2015). The composition of the gut microbiota throughout life, with an emphasis on early life. Microb. Ecol. Health Dis. *26*.

Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. *20*, 53–65.

Russo, G., Zegar, C., and Giordano, A. (2003). Advantages and limitations of microarray technology in human cancer. Oncogene *22*, 1206865.

Salgado, H., Moreno-Hagelsieb, G., Smith, T.F., and Collado-Vides, J. (2000). Operons in Escherichia coli: genomic analyses and predictions. Proc. Natl. Acad. Sci. U. S. A. *97*, 6652–6657.

Sanderson, K.E., and Roth, J.R. (1983). Linkage map of Salmonella typhimurium, Edition VI. Microbiol. Rev. *47*, 410–453.

Sandkvist, M. (2001). Type II Secretion and Pathogenesis. Infect. Immun. *69*, 3523–3535.

Sartor, R.B. (2008). Microbial Influences in Inflammatory Bowel Diseases. Gastroenterology *134*, 577–594.

Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995). Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. Science *270*, 467–470.

Schlegel, K., Welte, C., Deppenmeier, U., and Müller, V. (2012). Electron transport during aceticlastic methanogenesis by Methanosarcina acetivorans involves a sodium-translocating Rnf complex. FEBS J. *279*, 4444–4452.

Schmidt, H., and Hensel, M. (2004). Pathogenicity Islands in Bacterial Pathogenesis. Clin. Microbiol. Rev. *17*, 14–56.

Singer, H.M., Kühne, C., Deditius, J.A., Hughes, K.T., and Erhardt, M. (2014). The Salmonella Spi1 Virulence Regulatory Protein HilD Directly Activates Transcription of the Flagellar Master Operon flhDC. J. Bacteriol. *196*, 1448–1457.

Sinha, S., Cheng, S., Fan, C., and Bobik, T.A. (2012). The PduM Protein Is a Structural Component of the Microcompartments Involved in Coenzyme B12-Dependent 1,2-Propanediol Degradation by Salmonella enterica. J. Bacteriol. *194*, 1912–1918.

Spiga, L., Winter, M.G., Carvalho, T.F. de, Zhu, W., Hughes, E.R., Gillis, C.C., Behrendt, C.L., Kim, J., Chessa, D., Andrews-Polymenis, H.L., et al. (2017). An Oxidative Central Metabolism Enables Salmonella to Utilize Microbiota-Derived Succinate. Cell Host Microbe *22*, 291-301.e6.

Sridhar, J., Sekar, K., and Rafi, Z.A. (2009). CsrA interacting small RNAs in Haemophilus spp genomes: a theoretical analysis. Arch. Microbiol. *191*, 451–459.

Srikumar, S., and Fuchs, T.M. (2011). Ethanolamine Utilization Contributes to Proliferation of Salmonella enterica Serovar Typhimurium in Food and in Nematodes. Appl. Environ. Microbiol. *77*, 281–290.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. *102*, 15545–15550.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., et al. (2015). STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Res. *43*, D447–D452.

Tchawa Yimga, M., Leatham, M.P., Allen, J.H., Laux, D.C., Conway, T., and Cohen, P.S. (2006). Role of Gluconeogenesis and the Tricarboxylic Acid Cycle in the Virulence of Salmonella enterica Serovar Typhimurium in BALB/c Mice. Infect. Immun. *74*, 1130–1140.

Thaiss, C.A., Zmora, N., Levy, M., and Elinav, E. (2016). The microbiome and innate immunity. Nature *535*, 65–74.

Thiemann, S., Smit, N., Roy, U., Lesker, T.R., Gálvez, E.J.C., Helmecke, J., Basic, M., Bleich, A., Goodman, A.L., Kalinke, U., et al. (2017). Enhancement of IFNγ Production by Distinct Commensals Ameliorates Salmonella-Induced Disease. Cell Host Microbe *21*, 682-694.e5.

Thompson, C.C., Chimetto, L., Edwards, R.A., Swings, J., Stackebrandt, E., and Thompson, F.L. (2013). Microbial genomic taxonomy. BMC Genomics *14*, 913.

Trotochaud, A.E., and Wassarman, K.M. (2004). 6S RNA Function Enhances Long-Term Cell Survival. J. Bacteriol. *186*, 4978–4985.

Tu, X., Latifi, T., Bougdour, A., Gottesman, S., and Groisman, E.A. (2006). The PhoP/PhoQ two-component system stabilizes the alternative sigma factor RpoS in Salmonella enterica. Proc. Natl. Acad. Sci. *103*, 13503–13508.

Turnbaugh, P.J., and Gordon, J.I. (2009). The core gut microbiome, energy balance and obesity. J. Physiol. *587*, 4153–4158.

Vanderpool, C.K., and Gottesman, S. (2004). Involvement of a novel transcriptional activator and small RNA in post-transcriptional regulation of the glucose phosphoenolpyruvate phosphotransferase system. Mol. Microbiol. *54*, 1076–1089.

Vaz, J.B., Penha Filho, R.A.C., Junior, A.B., and Lemos, M.V.F. (2011). Requirement for cobalamin by Salmonella enterica serovars Typhimurium, Pullorum, Gallinarum and Enteritidis during infection in chickens. Braz. J. Microbiol. *42*, 1409–1418.

Wagner, G.P., Kin, K., and Lynch, V.J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci. *131*, 281–285.

Wells, J.N., Bergendahl, L.T., and Marsh, J.A. (2016). Operon Gene Order Is Optimized for Ordered Protein Complex Assembly. Cell Rep. *14*, 679–685.

Winter, S.E., Thiennimitr, P., Winter, M.G., Butler, B.P., Huseby, D.L., Crawford, R.W., Russell, J.M., Bevins, C.L., Adams, L.G., Tsolis, R.M., et al. (2010). Gut inflammation provides a respiratory electron acceptor for Salmonella. Nature *467*, 426–429.

Wu, H.-J., and Wu, E. (2012). The role of gut microbiota in immune homeostasis and autoimmunity. Gut Microbes *3*, 4–14.

Xiao, G., Martinez-Vaz, B., Pan, W., and Khodursky, A.B. (2006). Operon information improves gene expression estimation for cDNA microarrays. BMC Genomics *7*, 87.

Xiao, L., Feng, Q., Liang, S., Sonne, S.B., Xia, Z., Qiu, X., Li, X., Long, H., Zhang, J., Zhang, D., et al. (2015). A catalog of the mouse gut metagenome. Nat. Biotechnol. *33*, 1103–1108.

Xu, X., Xu, P., Ma, C., Tang, J., and Zhang, X. (2013). Gut microbiota, host health, and polysaccharides. Biotechnol. Adv. *31*, 318–337.

Yew, W.S., and Gerlt, J.A. (2002). Utilization of l-Ascorbate by Escherichia coli K-12: Assignments of Functions to Products of the yjf-sga and yia-sgb Operons. J. Bacteriol. *184*, 302–306.

Ze, X., Duncan, S.H., Louis, P., and Flint, H.J. (2012). Ruminococcus bromii is a keystone species for the degradation of resistant starch in the human colon. ISME J. *6*, 1535–1543.

Zeng, F., Wang, Z., Wang, Y., Zhou, J., and Chen, T. (2017). Large-scale 16S gene assembly using metagenomics shotgun sequences. Bioinformatics *33*, 1447–1456.

Zhang, Y.-Z., and Li, Y.-Y. (2014). Inflammatory bowel disease: Pathogenesis. World J. Gastroenterol. WJG *20*, 91–99.

Zhao, B., and Houry, W.A. (2010). Acid stress response in enteropathogenic gammaproteobacteria: an aptitude for survivalThis paper is one of a selection of papers published in this special issue entitled “Canadian Society of Biochemistry, Molecular & Cellular Biology 52nd Annual Meeting — Protein Folding: Principles and Diseases” and has undergone the Journal’s usual peer review process. Biochem. Cell Biol. *88*, 301–314.

ABILASH CHAKRAVARTHY DURAI RAJ

Schöppenstedterstraße 37 Phone: 004917670745041
Braunschweig Email:abilashchakravarthy@gmail.com

ACADEMIC QUALIFICATIONS

Qualification	Duration	Aggregate Percentage
Early Stage Researcher(PhD researcher)	2014-2018	
Masters in Life Science Informatics, University o Bonn	2011-2014	2.4(ECTS Scale)
Bachelors in Biotechnology, Anna University	2007-2011	85%
Higher Secondary School (Major : Biology, Maths, Physics, Chemistry) (TNHSC)	2006-2007	91%
Secondary School(TNHSC)	2004-2005	90%

RESEARCH EXPERIENCES

- ✓ **Doctoral study entitled "*In silico* modeling of bacterial transcriptional landscape in different intestinal microbial ecosystems during Salmonella infection" (August,2014-March,2018)**
  - ✓ Transcriptome-Metatranscriptome associative study to understand the potential Salmonella-gut microbiota interactions.
  - ✓ Develop Operon enrichment analysis for potential gene signatures identification.
  - ✓ Develop a scoring scheme *to integrate 16S and metagenome data*.
  - ✓ Network-based operon model to predict the novel target operons for transcription factors.
  - ✓ KEGG Orthology-based Salmonella -gut microbiota network to identify potential microbial cross-talk and their associations with 16S data.

- ✓ **Research Assistant worked in "Constraint-based modeling in metabolic networks" (May,2014- August,2014)**
  - ✓ Worked in University of Düsseldorf Bioinformatics Department on identification and using blocked reactions from various Ecooli strains for phylogeny studies.
- ✓ **Research Assistant worked in "Comparative analysis of pancreatic tumour therapeutics using integrative omics data approaches"(December,2013 – May,2014)**
  - ✓ Worked in LIMES Institute, University of Bonn to **integrate metabolomics and micro-array data** from pancreatic tumour patients to study therapeutic effects.
- ✓ **Master's Thesis in "Network-based biomarkers in metabolic systems: algorithms and applications" (April,2013 – November,2013)**
  - ✓ Master's Thesis carried out in the **ICB Department, Helmholtz Zentrum, Munich** where the project is specifically aimed at developing and extending graph clustering algorithms to obtain potential network based biomarkers from the metabolome data. The effect of lowered or raised reaction rates in all extreme pathways involved in a simulated reaction system is observed and studied.
  - ✓ **Bio-Imaging (April,2013-November,2013)**
    - Worked as a "Student Assistant" on Cell Tracking using TTT software and Segmentation using FIJI at the Helmholtz Zentrum, Munich.
- ✓ **Literature Mining (October 2012-January 2013)**
  - ✓ Worked in the Fraunhofer Institute SCAI in text mining in a NeuroAllianz Project as a student assistant where Protein-Protein Interaction Networks were obtained from the literature for Alzheimer's disease. Work was also carried out in developing and extending ontology for Neuro-Degenerative Diseases.

TECHNICAL KNOWLEDGE

- ✓ **Computer languages**: R & Cluster computing (Bash), Matlab, Python, SQL,JAVA, C, C++.
- ✓ Handled Sequence data ranging from transcriptomics, genomics/metagenomics data to 16S data.
- ✓ Hands on experience on Mathematical modeling, Data Visualization techniques and Machine learning approaches.
- ✓ Worked in close association with the experimental biologists in generating and validating hypothesis.

- ✓ Obtained excellent grades, in-depth knowledge and hands-on experience in molecular biology, cell biology during undergraduate study.
- ✓ **Languages**: English (Fluent), Tamil (Native), German (A2) and Hindi (Beginner).

FIELDS OF INTEREST

Systems Biology, Genomics, Reverse Ecology, Network Biology, Probabilistic graphical modeling, Big data analysis, Bio-statistics, Molecular Biology, Microbiology.

PUBLICATIONS

- ✓ [Microbiota-dependent reprogramming of the transcriptional profile of Salmonella Typhimurium] **A.Chakravarthy**, S. Thiemann, E. Gálvez, T. Lesker, N. Smit, K. Hokamp, C. Kröger, T. Strowig, M. Erhardt
  in process

- ✓ [A integrated mouse gut gene catalogue with high resolution taxonomic profiling by linking 16S and metagemic species reconstruction]
  T. Lesker, E. Gálvez, **A. Chakravarthy**, J. Baines, T. Clavel, A. Sczyrba, A. McHardy, T. Strowig
  in process
    - Multi-omics integration (16S to MGS Bins), comparative analysis of mouse-gut catalogues, associative study between diet intervention and sub-microbial communities.

PERSONAL STRENGTHS

I am very sincere, dedicated, hardworking and a good team-player. Inquisitiveness coupled with perseverance defines my personality. I am a self-motivated person, who gives that extra push to give my best and have an urge to think higher and reach my goal. My career interest is in becoming a Research Scientist in Bioinformatics.

References

Dr. Till Strowig	**Prof. Dr. Marc Erhardt**	**Prof. Fabian Theis**
Junior Research Group Leader	Group Leader	Head of the Institute
Microbial Immune Regulation	Bacterial Physiology	Institute of Computational Biology
Helmholtz Center for Infectio Research, Braunschweig, Germany.	Humboldt-Universität zu Berli Germany.	Helmholtz Zentrum, Munich,Germany.
Till.Strowig@helmholtz-hzi.de	marc.erhardt@hu-berlin.de	fabian.theis@helmholtz-muenchen.de